AN ADAPTIVE SPECTRUM ANALYSIS VOCODER

A THESIS

Presented to

The Faculty of the Graduate Division

by

Jack Curtis Hammett, Jr.

In Partial Fulfillment

of the Requirements for the Degree

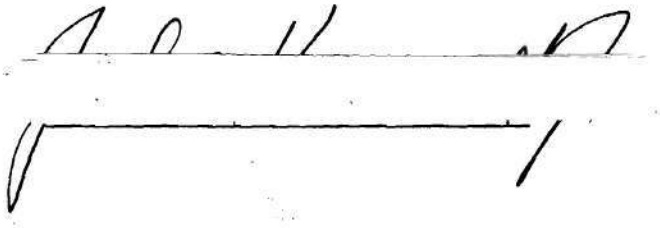
Doctor of Philosophy

in the School of Electrical Engineering

Georgia Institute of Technology

May, 1971

In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



7/25/68

AN ADAPTIVE SPECTRUM ANALYSIS VOCODER

Approved:

Date approved by Chairman: May 24, 1971

ACKNOWLEDGMENTS

The technical leadership and day-to-day encouragement of Dr. Aubrey M. Bush, my thesis advisor, motivated the conception and execution of this project.

Dr. Benjamin J. Dasher served on the proposal and reading committees, and provided many stimulating discussions about speech communications.

Dr. Ronal W. Larson served as chairman of the proposal committee and as a member of the reading committee, and offered many useful suggestions.

Mr. Alton P. Jensen, Senior Research Engineer in the School of Information and Computer Science, wrote the first assembly language program for real-time digital-to-analog conversion of speech on the DEC PDP-8 and provided guidance on small computer operations.

Dr. Vladimir Slamecka, Director of the School of Information and Computer Science, made the PDP-8 facility available to me on a routine basis.

Mr. Douglas W. Robertson, Head of the Communications Branch of the Electronics Division, Engineering Experiment Station, made his facilities available to me, and displayed continuing interest in the speech research.

The many employees of the Rich Electronic Computer Center gave me valuable assistance in completing the UNIVAC 1108 portion of the effort.

Carol, my wife, assembled the spectrum models and shared with me the life of a graduate student.

To these people, I express my sincere appreciation and thanks.

While performing this work, I was supported by the United States Army through the Advanced Degree Program.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Concluded)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Concluded)

SUMMARY

An adaptive spectrum analysis strategy was incorporated into the short-time spectrum coding of vocal tract information in the vocoder. The vocoder is based on a simplified model of speech production in which the speech signal is viewed as the response of the vocal tract "filter" to either a periodic or a noise-like source. The vocal tract spectrum information is obtained by smoothing the short-time spectrum surface. Samples of the surface are coded for transmission in a reduced capacity channel. An approximation to the original speech is synthesized by exciting a filter which is controlled by the received spectrum parameters. The vocoder attempts to preserve the shape of the short-time spectrum.

The simplified model is an essentially "stationary" one, valid to the extent that the input speech is stationary in each analysis-synthesis interval. Modern vocoders employ a fixed time-frequency compromise determined by the duration of the window function used in short-time spectrum analysis, the degree of smoothing of the spectrum surface, and the pattern of sampling of the spectrum for transmission.

The phonemes of speech display a wide range of time-frequency properties, due to the extremes in the articulatory dynamics of speech production. The time-frequency compromise in the short-time spectrum was examined with respect to the stationary model and the dynamics of speech production. The conclusion reached is that improvement in vocoded speech "quality" might be achieved by adapting the (time-frequency) resolution

cell in the analyzer to the nature of the segment of speech in each analysis frame.

The adaptive strategy was incorporated into the design of a homomorphic vocoder simulation. Since the homomorphic vocoder codes the vocal tract spectrum as the low-time parameters of the cepstrum, the frequency resolution retained in the coding depends on the number of parameters transmitted. Time resolution is readily manipulated by adjusting the duration of the analysis window. The adaptive action is controlled by "frame decisions" which were made by hand, based on a cepstrum distance criterion, after a preliminary analysis which also served the pitch detection function.

Experimental runs with the adaptive homomorphic vocoder were made with three test sentences and the synthesized speech was judged in informal subjective listening tests. In one experiment (with a female talker) two adaptive modes were employed with window durations of 12.8 or 25.6 ms, frame intervals of 10 or 20 ms, and cepstrum truncation to 10 or 20 coefficients, respectively. The spectrum data rate was reduced to 3700 b/s and the synthesized speech judged to be of high "quality," retaining naturalness and recognition properties.

Two additional experiments (with male talkers) used window durations of 10 or 20 ms and a 3700 b/s data rate. The first of these resulted in synthesized speech judged to be of high "quality" but slightly less natural than the earlier result. The last experiment was conducted with a test sentence composed of voiced, non-nasal phonemes, which displayed no transitions in the spectrum rapid enough to warrant use of the

10 ms window mode, so the simulation operated as a conventional homomorphic vocoder with a 20 ms window. The result was judged to be reasonably good, but relatively not quite as good as the two previous results.

The tentative conclusion of the experimental phase of the investigation is that the adaptive strategy has potential for reducing vocoder data rates, while maintaining intelligibility, speaker recognition, and naturalness properties.

CHAPTER I

INTRODUCTION

The objective of this research is to improve the performance of modern speech bandwidth compression systems. The properties of the vocoder (voice-coder) are examined, a potential improvement is proposed and incorporated into a vocoder design, and the performance of the resulting system is evaluated by computer simulation.

The speech analysis and synthesis strategy of the vocoder is examined in Chapter II. The strategy is to extract from the speech signal slowly-varying parameters which describe the physical configuration of the vocal tract and the nature of its excitation. These parameters, obtained by analysis, are used to synthesize an approximation to the original speech. The strategy is based on the mechanics of human speech production. The simplified model of speech production motivates short-time spectrum analysis, the technique used in the channel vocoder, the formant vocoder, and the homomorphic vocoder.

The channel vocoder employs a bank of band-pass filters with center frequencies spaced across the speech band. The smoothed filter outputs, which represent the vocal tract spectrum, are slowly-varying signals which are transmitted in a reduced bandwidth channel. In a similar fashion, the formant vocoder tracks the regions of high energy in the spectrum, and transmits formant frequency and amplitude parameters.

Speech synthesis is accomplished in both channel and formant vocoders by exciting a filter whose system function is controlled by the received spectrum signals.

The homomorphic vocoder performs a deconvolution of the speech waveform by linear filtering the log spectrum. The low-order parameters of the cepstrum, a Fourier expansion of the log spectrum, are the transmitted spectrum signals. The cepstrum parameters are transformed into a vocal tract impulse response function, which is convolved with an excitation signal to produce synthesized speech.

The objective of Chapter III is to examine the time-frequency compromise inherent in the short-time spectrum analysis employed in vocoder systems. The uncertainty principle and the scaling property of Fourier analysis, and the concept of dynamic "bandwidth" and "duration," are related to the role of the window function in short-time spectrum analysis. The notion of a resolution rectangle which describes the time and frequency properties of a spectrum analyzer is employed to study the compromise in vocoder systems. The conclusion reached is that an adaptive strategy of adjusting the resolution cell on a short-time basis may improve the "quality" of vocoded speech.

The design of an adaptive homomorphic vocoder system is described in Chapter IV. A frame-decision strategy is selected which is keyed to a distance measure in the short-time cepstrum. The frame decision determines which adaptive "mode" is to be used to analyze the "present" frame. Thus, the window function duration and the number of transmitted cepstrum coefficients are adjusted on a frame-by-frame basis. The adaptive voco-

der design is described, and the experimental simulation system outlined.

A description of the experiments performed and the results obtained is given in Chapter V. The conventional and adaptive homomorphic vocoders were simulated in various configurations, and the "quality" of the resulting synthesized speech judged in informal subjective listening tests. The results obtained with the adaptive vocoder suggest that the adaptive scheme has potential for reducing the required spectrum data rate while retaining intelligibility, naturalness, and speaker recognition properties. The spectrum bit rate was reduced from 6800 b/s to 3700 b/s without noticeable degradation in "quality."

The tentative conclusion of the experimental study is that the adaptive approach permits substantial reduction in vocoder bit rates without loss of "quality."

Recommendations for further work are outlined in Chapter VI. Six areas of further investigation are proposed.

CHAPTER II

SPEECH ANALYSIS AND SYNTHESIS

The goal of speech analysis and synthesis is to obtain an efficient representation of speech information. Such a representation permits communication over channels of smaller capacity and storage of larger quantities of speech in a digital memory for voice response. Another application is time expansion-compression of speech. Time expanded speech may be useful in language training and speech therapy, while time compressed speech has application to talking books for the blind and to computer-aided instruction. Modern analysis-synthesis systems represent speech in data rates of 1000 to 8400 bits/second (b/s) rather than the 50,000 b/s required to represent the acoustic waveform [1,2].

Introduction

Radio and telephone communications systems convey a speech signal by preservation of the acoustic waveform. A substantial mismatch exists between the capacity of the human to generate and perceive information and the capacity of the "waveform" channel. The channel is capable of much higher information rates than is the human. For conversational speech only about 50 bits/second would be required to transmit the written equivalent. Conversely, a typical voice channel with 3 kHz bandwidth and 30 db signal to noise ratio has an information capacity on the order of 30,000 bits/second [1]. Of course, the acoustic waveform contains more

information than the written equivalent, but certainly not 600 times more. The acoustic waveform is clearly not an efficient code for speech information.

The objective of this research is to improve the performance of modern speech bandwidth compression systems. A technique of adaptive spectrum analysis is proposed and incorporated into a vocoder (voice-coder) design.

Speech Production [1]

Human speech originates in the larynx -- a box-like structure of cartilage at the upper end of the trachea. The larynx houses two lips of ligament and muscle called the vocal cords. The opening between the vocal cords is called the glottis. A drawing of the human vocal mechanism is shown in Figure 1.

Voiced speech is produced by forcing air through the glottis while the vocal cords are held under tension. The glottis vibrates open and shut generating a quasi-periodic flow of air -- a pulse train acoustic time function rich in harmonics. The fundamental frequency of the vocal cord oscillation is called the voice pitch.

The excitation signal from the glottis passes through the vocal tract, which includes the throat, mouth, and nasal cavity. The message the talker wants to convey is imposed on the excitation signal by the changes in position of the tongue, lips, and other moving parts of the tract. These moving parts are called articulators, and their activity in creating the spoken language is called articulation. During articulation the vocal cavity assumes different positions causing resonances in the

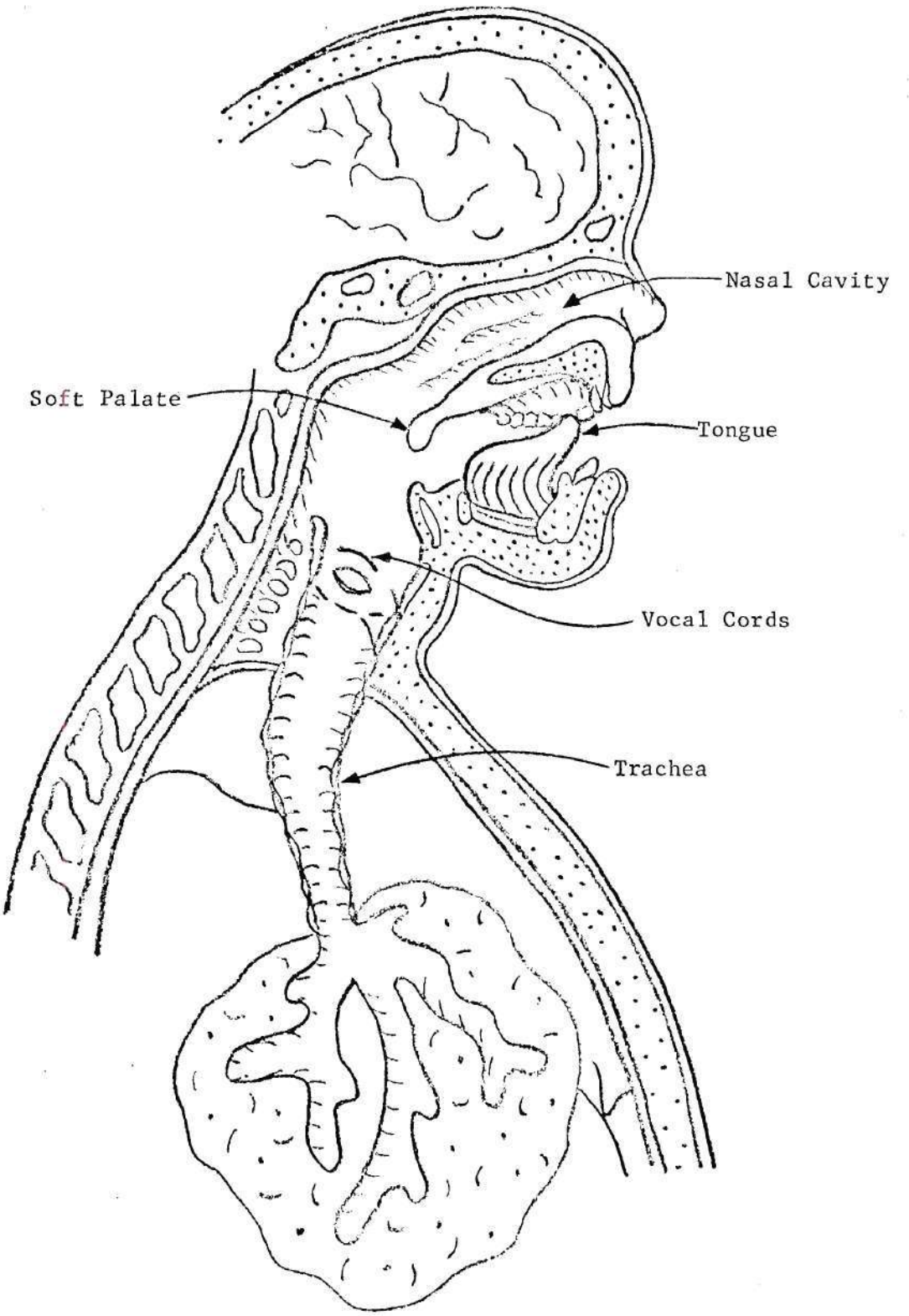


Figure 1. The Human Vocal Mechanism

tract which alter the spectrum of the excitation signal, imposing on the spectrum peaks which are called formants.

Unvoiced speech is produced by a turbulent flow of air past a constriction in the vocal tract, or by a release of pressure at some point of closure in the tract. Unvoiced excitation is an acoustic noise source. The spectrum of an unvoiced speech sound is influenced mainly by that portion of the vocal tract forward of the constriction. Pressure released at a closure causes an initial burst, followed by turbulent flow noise.

The phoneme is the smallest unit of speech that distinguishes one utterance from another. General American English has about 42 phonemes [1]. We may think of these phonemes as a code uniquely related to the articulatory gestures of the language.

The vowel sounds of speech are produced by voiced excitation of the vocal tract (e.g., the "ah" in father). In normal articulation the tract is held in a relatively stable position during most of the sound. Vowels usually have a "duration" of 60 ms or longer. The soft palate seals off the passageway between the oral cavity and the nasal cavity during vowel production. The vowel phonemes are listed below.

Vowels

i (EVE)	I (IT)	e (HATE)
ɛ (MET)	æ (AT)	ɑ (FATHER)
ɔ (ALL)	o (OBEY)	ʊ (FOOT)
u (BOOT)	ʌ (UP)	ɝ (BIRD)

Aside from the vowels, the remaining phonemes are referred to as consonants, some of which are more transitory in nature.

Fricative consonants are produced by incoherent noise excitation of the vocal tract caused by turbulent air flow at a constriction (e.g., the "s" in see). Radiation of fricatives occurs from the mouth. The vocal cord source may operate in conjunction with the noise source to produce a voiced fricative (e.g., the "z" in zoo).

Fricative Consonants

Voiced	Unvoiced
v (VOTE)	f (FOR)
ð (THEN)	θ (THIN)
z (ZOO)	s (SEE)
ʒ (AZURE)	ʃ (SHE)
	h (HE)

Stop consonants are produced by the abrupt release of pressure at a place of closure in the tract (e.g., the "t" in to). The articulatory movements which generate stops are more rapid than for other sounds. Stops may be voiced or unvoiced.

Stop Consonants

Voiced	Unvoiced
b (BE)	p (PAY)
d (DAY)	t (TO)
g (GO)	k (KEY)

The nasal consonants are voiced sounds characterized by a complete closure toward the front of the oral cavity -- at the lips or by the tongue. The soft palate is open wide so the nasal tract is the transmission channel.

Nasal Consonants

m (ME)

n (NO)

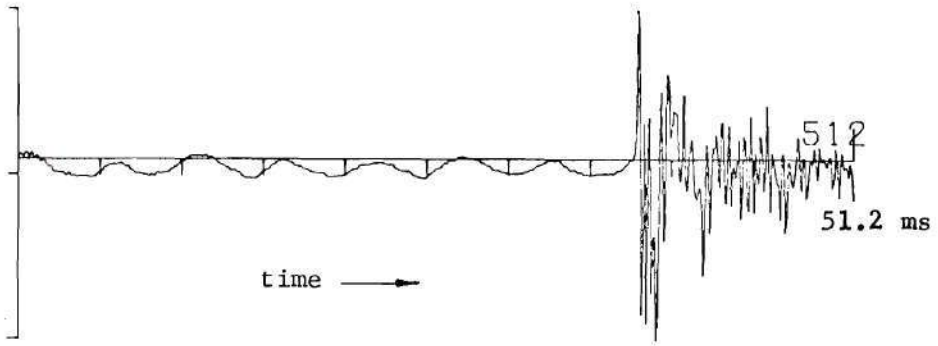
ŋ (SING)

The remaining consonants are classified as glides, semivowels, diphthongs, and affricates.

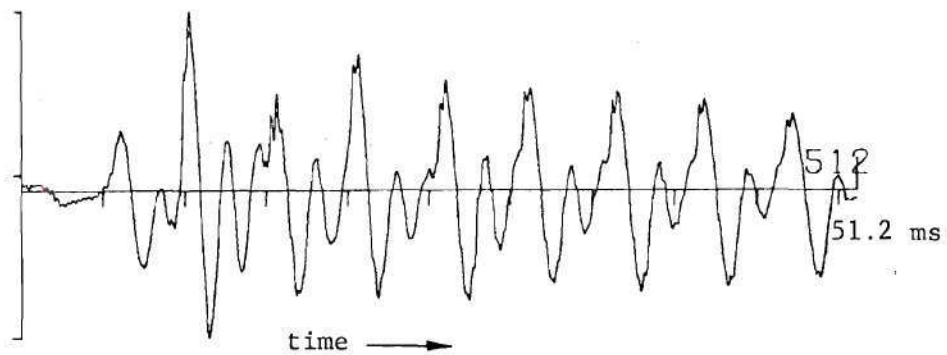
Other Consonants

Glides	Semivowels	Diphthongs	Affricates
w (WE)	r (READ)	eI (SAY)	tʃ (CHEW)
j (YOU)	l (LET)	Iu (NEW)	dʒ (JAR)
		ɔI (BOY)	
		aU (OUT)	
		aI (I)	
		oU (GO)	

The waveforms of the two phonemes /t/ and /I/ (spoken by a female talker) are shown in Figure 2. Each plot illustrates the normalized waveform of a 51.2 ms segment of conversational speech. The scale on the abscissa is 12 ms/in. In Part (a) is shown the stop consonant /t/ in the utterance "gift is." The closure portion of the /t/ lasts about 40 ms. The silence of closure is followed by a burst, signaling the release of pressure at the closure. The burst response decays with roughly exponential envelope to a low level aspirated noise. (To aspirate is to pronounce with a breathing, as in the fricative /h/.) The burst decays in 10 ms, while the following aspiration lasts about 20 ms before the buildup of the vowel /I/. A 51.2 ms segment of the connected vowel /I/ is shown in Part (b). The last 5 ms of the aspirated noise is seen at the extreme



(a)



(b)

Figure 2. Speech Waveforms (a) The Stop Consonant /t/ in "gift is," (b) The Vowel /I/ in "gift is"

left. Voicing begins, with a pitch of approximately 185 Hz, and within about 15 ms the waveform displays the steady state character of a sustained vowel. The maximum amplitude of the vowel in Part (b) is five times that of the stop consonant in Part (a).

A key observation about the phonemes of speech is that some sounds (e.g., the stop consonants) are produced by a rapid motion of the articulators. The vocal dynamics of such sounds are manifest in the acoustic waveform, which displays prominent short-time features. On the other hand, the vowels and the consonants which are continuants may be uttered as sustained sounds, requiring no articulatory motion. The waveforms of such sustained sounds display a qualitative quasi-periodic or quasi-stationary character. The striking difference between the character of "short" sounds and "long" sounds is central to the theme of this thesis. In the following pages we will examine this difference from many viewpoints, and seek insight into its importance in speech analysis-synthesis systems.

The Simplified Model of Speech Production

The speech waveform may be decomposed into excitation and vocal tract components. The complexity of the speech waveform is due principally to the broadband nature of the excitation signal. But the excitation signal may be adequately described by its pitch during voiced sounds. The vocal tract is a simple acoustic cavity which may be described by parameters which vary slowly -- at syllabic rates. The simplified descriptions of excitation and vocal tract may be transmitted over a reduced capacity communication channel and the two components reconstituted into

an intelligible, high-quality synthetic version of the original speech signal.

Decomposition into excitation and vocal tract components is the central strategy of nearly every speech analysis-synthesis system. The strategy was conceived by Homer Dudley at the Bell Telephone Laboratories and incorporated into his invention of the channel vocoder in 1939 [3]. Speech bandwidth reduction (and data reduction) research during the last three decades has been dominated by the intuitive strategy pioneered by Dudley.

Speech may be viewed as the response of a linear system (the vocal tract "filter") to one or more sound sources. This statement is the essence of the acoustic theory of speech production described by Gunnar Fant [4]. During a voiced sound, a periodic pulse train excitation signal $e(t)$ is the input to the linear time-invariant vocal tract "filter" which has impulse response $v(t)$. The resulting speech waveform is $s(t) = e(t) \circledast v(t)$. Unvoiced sounds are assumed to be produced in the same fashion, except that $e(t)$ is considered to have the character of stationary random-noise.

We notice that the word "stationary" is appropriate to describe each of the three properties of the simplified model:

1. $v(t)$ is the impulse response of a time-invariant filter, so the vocal tract is considered to be "stationary" in the sense that the articulators are not moving.

2. For unvoiced sounds, $e(t)$ is modeled as a stationary random process, so the excitation is "stationary" in the sense that the statistics of the random process are unchanged by a shift in the time origin.

3. For voiced sounds, e(t) is modeled as a periodic signal, so the excitation is "stationary" in the sense that its description does not change with time.

Thus, we will refer to the simplified model as the stationary model to emphasize its time-invariant character. The stationary model of speech production is illustrated in Figure 3.

The stationary model is an adequate representation of sustained vowel and continuant phoneme production, under steady state conditions. Some approximation is involved, however, since voice pitch often displays a slight "jitter" -- so the excitation is not exactly periodic in the short-term sense, nor is the vocal cavity ideally linear. But the approximation is a good one. The weakness of the stationary model obtains from its time-invariant simplicity.

Since the articulatory gestures in speech production are relatively slow -- compared, say to a pitch period -- we may think of the speech waveform as being constructed of short segments, each segment corresponding to a fixed vocal tract configuration. The motion of the articulators is continuous, so the stationary model may be used for successive short segments during which only incremental movement of the tract occurs. This approach to the analysis and subsequent synthesis of speech is clearly valid to the extent that excitation and vocal tract are stationary in each segment. In Chapter III we will consider further the validity of the stationary model.

Decomposing speech into excitation and vocal tract components and the use of the stationary assumption are the key features of that class of analysis-synthesis systems known as vocoders. Many vocoder realizations

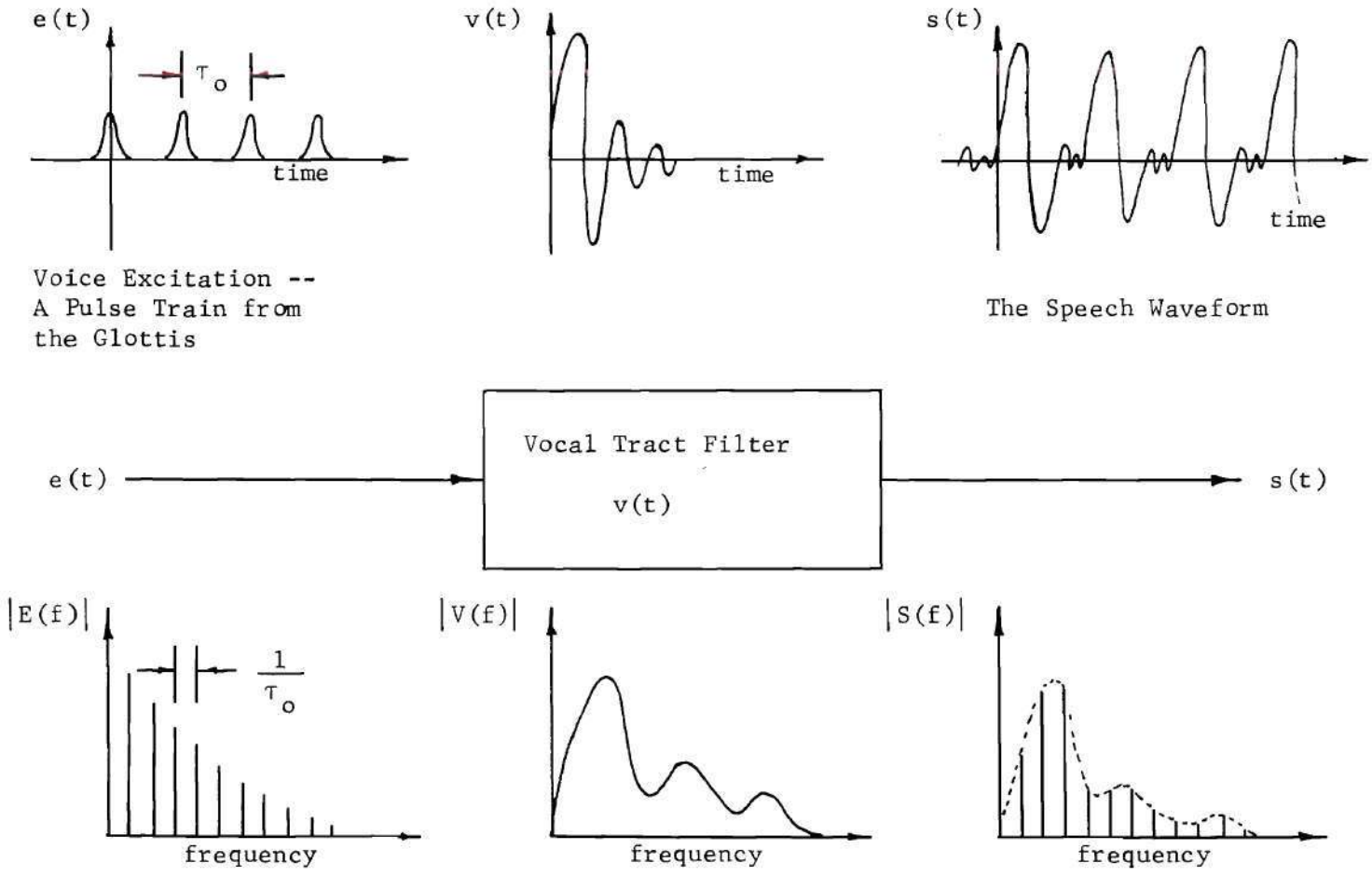


Figure 3. The Stationary Model of Speech Production

have been invented, but all employ a coding of excitation and vocal tract components and a short-term stationary synthesis strategy. A vocoder system diagram is shown in Figure 4.

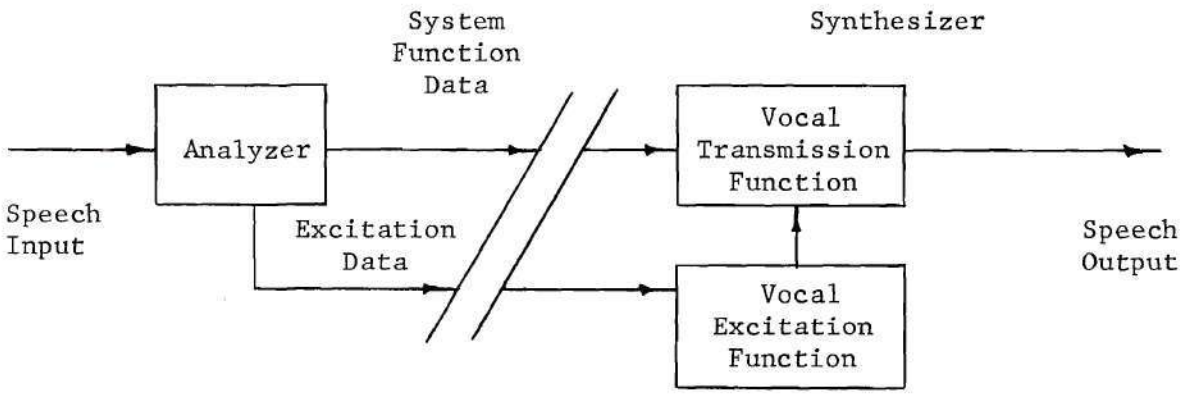


Figure 4. The Vocoder

Another description of the strategy of the vocoder is that it attempts to preserve the short-time spectrum of the speech signal. We examine the short-time spectrum in the next section.

Spectrum Analysis of Speech

The traditional tool for spectrum analysis of signals and linear systems is the Fourier transform pair

$$S(f) = \int_{-\infty}^{\infty} s(t)\, e^{-j2\pi ft}\, dt \qquad s(t) = \int_{-\infty}^{\infty} S(f)\, e^{+j2\pi ft}\, df \qquad (2\text{-}1)$$

which we write symbolically as

$$s(t) \underset{t \quad f}{\longleftrightarrow} S(f)$$

A speech signal, however, neither satisfies the condition

$$\int_{-\infty}^{\infty} |s(t)|\, dt < \infty$$

nor is known over all time. We may modify (2-1) so the integration extends only over the known past of s(t):

$$S_t(f) = \int_{-\infty}^{t} s(\tau)\, e^{-j2\pi f\tau}\, d\tau \qquad (2\text{-}2)$$

$$s(\tau)\, U_{-1}\, (t - \tau) \underset{\tau \quad f}{\longleftrightarrow} S_t(f)$$

where $U_{-1}(t)$ is the unit step function. $S_t(f)$ is known as the running spectrum of s(t) [5]. $S_t(f)$ corresponds to the Fourier transform of that portion of the time function s(t) that is "seen" through a step function "window" which obscures the future.

Short-Time Spectrum Analysis

For speech we desire a frequency representation which puts in evidence the spectral content of short segments of speech. To this end, we choose a window function w(t) which is causal and essentially non-zero only over a duration D. Forming the product

$$s(\tau)\, w\, (t - \tau)$$

and Fourier transforming yields the short-time spectrum S(t, f) [1]:

$$S(t, f) = \int_{-\infty}^{t} s(\tau)\, w\, (t - \tau)\, e^{-j2\pi f\tau}\, d\tau \qquad (2\text{-}3)$$

$$s(\tau)\ w\ (t - \tau) \underset{\tau}{\overset{f}{\longleftrightarrow}} S(t,\ f)$$

The short-time spectrum $S(t, f)$ is the Fourier transform of the recent past of the time function $s(\tau)$ weighted by the window function $w(t - \tau)$. An illustration of the role of the window function is shown in Figure 5.

We may interpret equation (2-3) with the aid of Figure 5. The window function $w(\tau)$ reversed in time, $w(t - \tau)$, moves (to the right) along the time axis, with its leading edge at running time t. After multiplication by the input speech signal $s(\tau)$, the product $s(\tau)\ w(t - \tau)$ is Fourier transformed with respect to the time variable τ. This process is performed at each successive running time instant, t, providing a continuum of spectrums, $S(t, f)$, which describe a (complex) surface in the (complex) intensity dimension, S, above the time-frequency plane.

By changing the variable τ in equation (2-3) to $t - \tau$, we obtain an equivalent definition for the short-time spectrum:

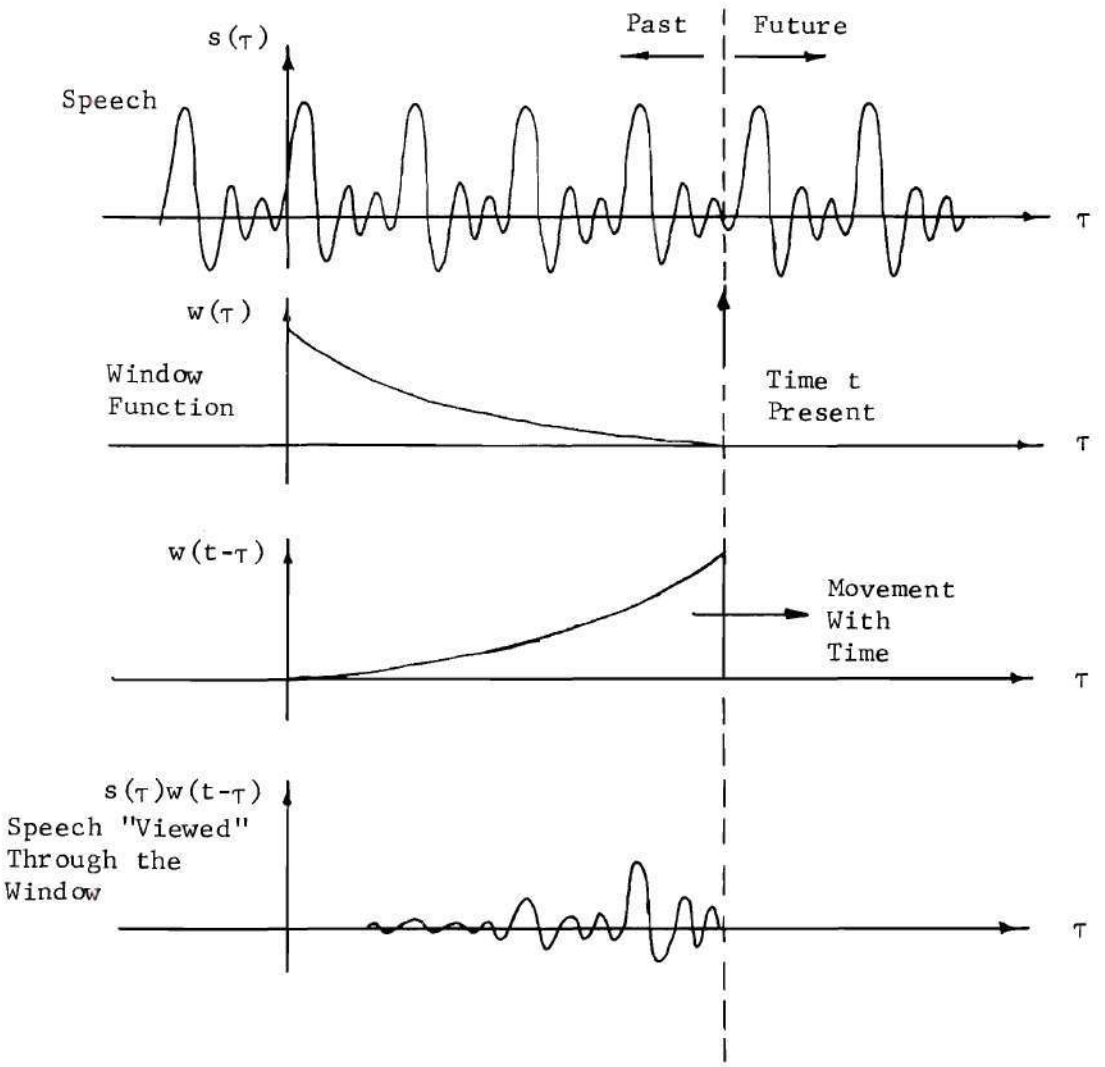
$$S(t,\ f) = e^{-j2\pi ft} \int_0^\infty s(t - \tau)\ w(\tau)\ e^{+j2\pi f\tau}\ d\tau \qquad (2\text{-}4)$$

Using the Euler identity in (2-4) we obtain

$$S(t,\ f) = e^{-j2\pi ft} \left\{ \int_0^\infty s(t - \tau)\ w(\tau)\ \cos\ 2\pi f\tau\ d\tau \right. \qquad (2\text{-}5)$$

$$\left. + j \int_0^\infty s(t - \tau)\ w(\tau)\ \sin\ 2\pi f\tau\ d\tau \right\}$$

$$= e^{-j2\pi ft} \left\{ a(t,\ f) + j\ b(t,\ f) \right\}$$



$$S(t,f) = \int_{-\infty}^{t} s(t)\, w(t-\tau) e^{-j2\pi f\tau}\, d\tau$$

Figure 5. The Short-Time Spectrum

The integrals a(t, f) and b(t, f) in (2-5) are the convolution of s(t) with w(t) cos $2\pi ft$ and w(t) sin $2\pi ft$, respectively. Thus, we may generate the short-time amplitude spectrum,

$$|S(t, f_1)| = [a^2(t, f_1) + b^2(t, f_1)]^{1/2}$$

by quadrature filtering. In practice an approximate version of $|S(t, f_1)|$ is obtained by generating the time envelope of either $a(t, f_1)$ or $b(t, f_1)$ [1]. Generation of the short-time amplitude spectrum is illustrated in Figure 6. Each system shown produces at its output a time function which describes the short-time spectrum surface at one frequency, f_1, in the time-frequency plane. By employing a bank of such filter systems with center frequencies $\{f_i\}$ spaced across the speech band, we obtain a coding of the short-time spectrum as the set of channel signals $\{|S(t, f_i)|\}$. Such a coding is used in the channel vocoder, described in the next section.

The systems shown in Figure 6 represent realizations of equation (2-5). An equivalent pair of systems follows from equation (2-3). The band-pass filters in Figure 6 are replaced by a multiplier-filter; s(t) is multiplied by cos $2\pi f_1 t$ (or sin $2\pi f_1 t$) and filtered with impulse response w(t).

Two conceptually dual models for generating the short-time spectrum are shown in Figure 7. Part (a) illustrates a system which describes the short-time spectrum as discussed above. The (complex) surface is described at one frequency, f_i, by the (complex) time function $S(t, f_i)$. To complete

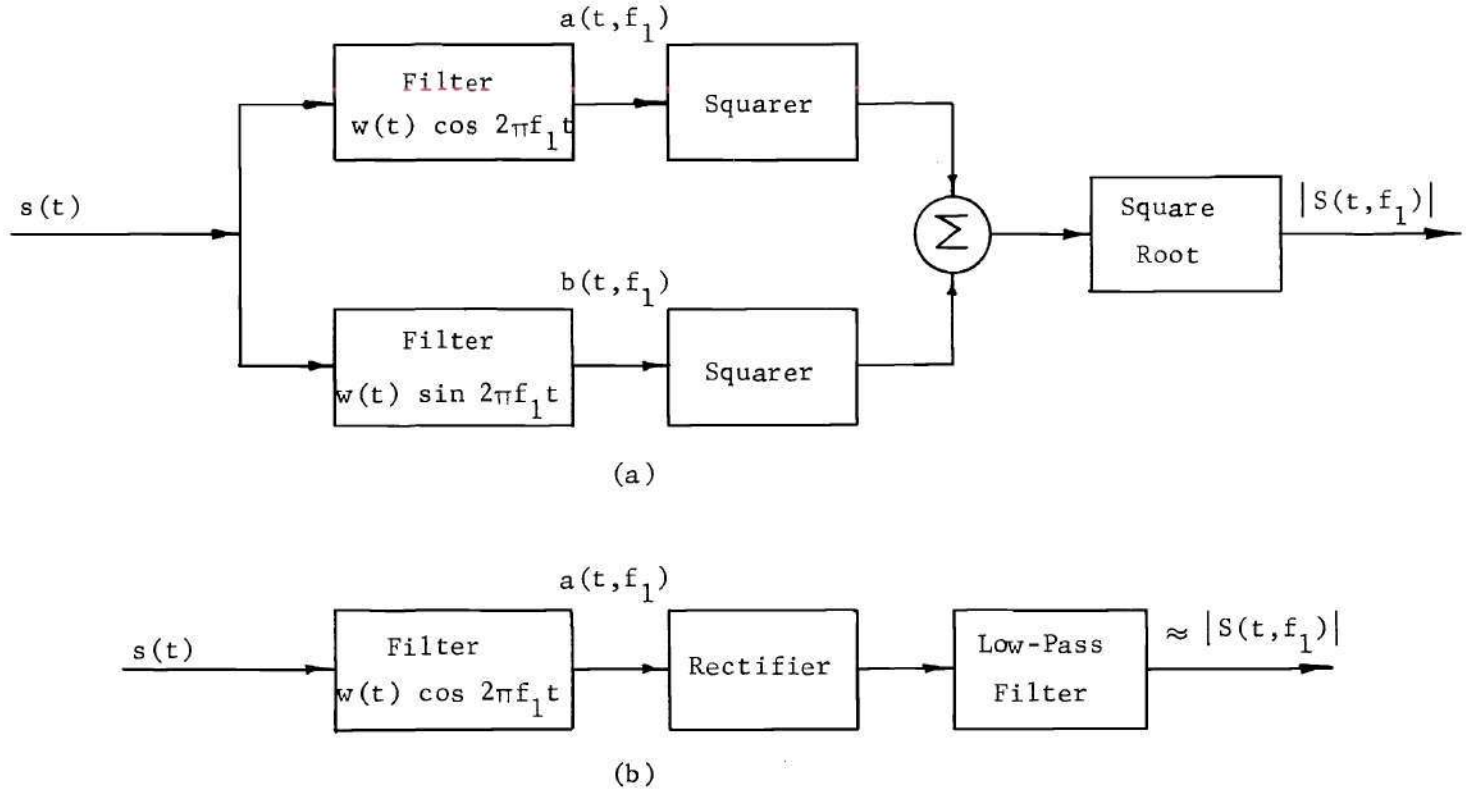


Figure 6. Generation of the Short-Time Spectrum
(a) Quadrature Filtering -- A Time Section at Frequency f_1
(b) An Approximate Method

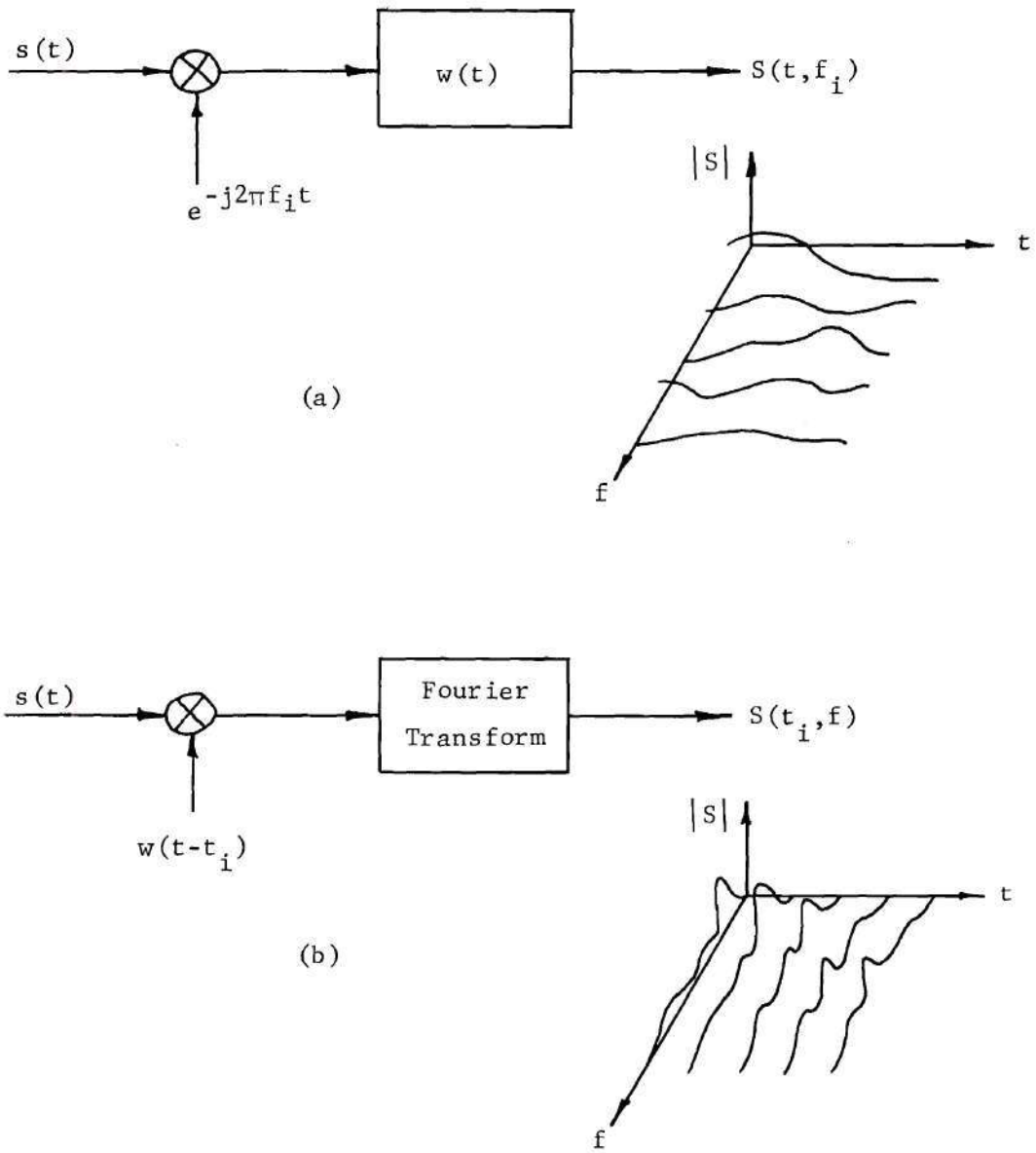


Figure 7. Dual Models of Short-Time Spectrum Generation
(a) Analysis at Discrete Frequencies -- Time Sections
(b) Analysis at Discrete Times -- Frequency Sections

the description of the surface a set of time sections is obtained at frequencies $\{f_i\}$. Part (b) illustrates the dual viewpoint. One section of the (complex) surface is described at time t_i by the (complex) frequency function $S(t_i, f)$. The description is completed by obtaining a set of frequency sections at times $\{t_i\}$.

Analysis of speech at discrete frequencies (as in Part (a) of Figure 7) is the approach employed in analog spectrum analyzers, whereas both viewpoints are commonly used in digital processors. Analysis at discrete times (as in Part (b)) became a practical technique when the Fast Fourier Transform algorithm was reported in 1965 [6].

A sketch of the short-time amplitude spectrum surface for a simple signal is shown in Figure 8. The signal is the sum of two similar components whose energy is concentrated in different regions of the time-frequency plane. Each component has a decaying exponential envelope and is modulated by a complex exponential. An exponential window was used in the short-time spectrum calculation.

The short-time spectrum describes the distribution of energy in frequency as it changes with time. The generation and coding of this short-time spectrum are key to most speech data-rate reduction systems.

The choice of the window function is crucial in the design of a spectrum analyzer. The scaling property of Fourier transforms shows that choice of a window function involves a necessary compromise between the time "resolution" and frequency "resolution" that may be achieved in the short-time spectrum. Another consideration is that the duration of the window function bears a vital relation to the validity of the stationary assumption. We will examine these issues in detail in Chapter III.

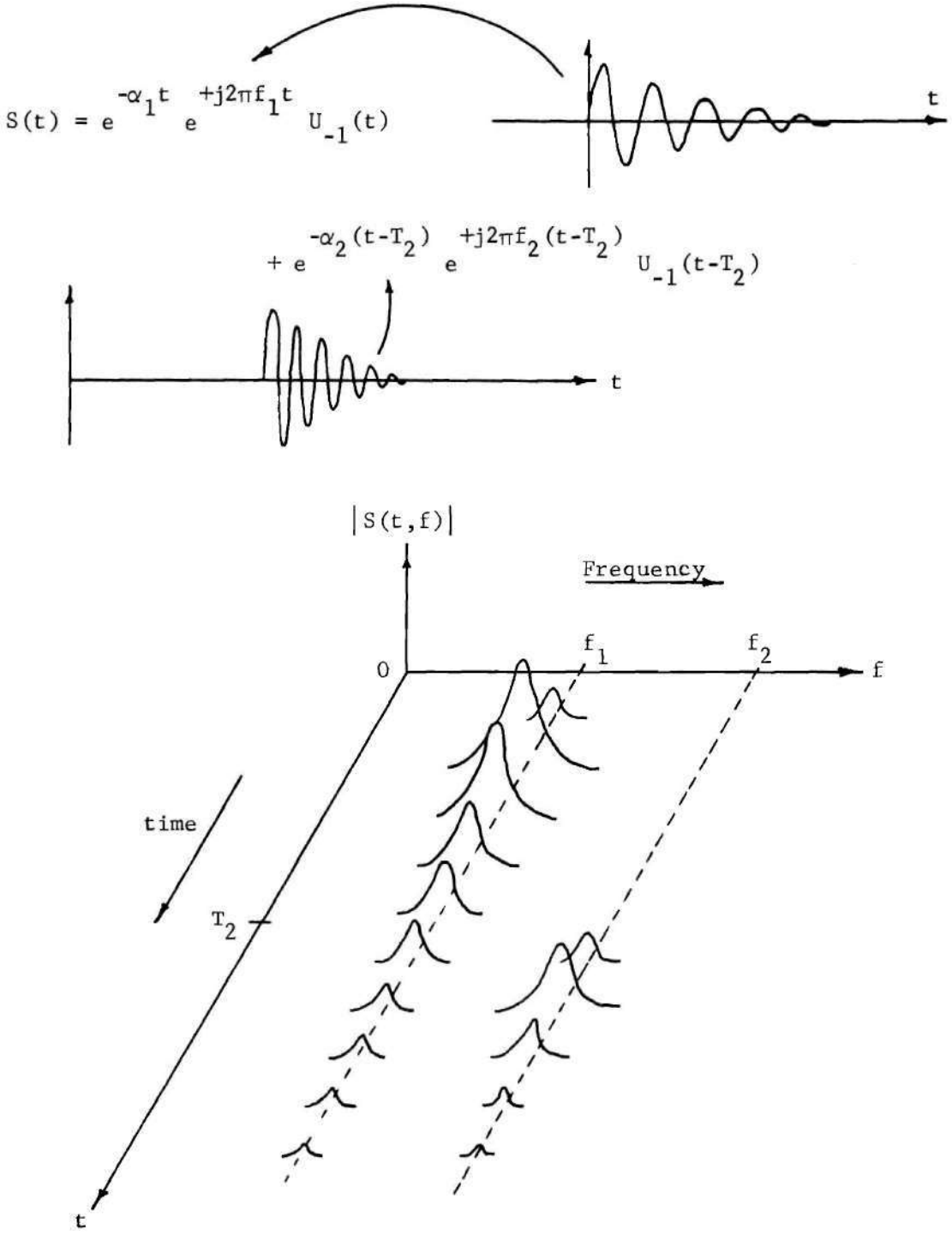


Figure 8. The Short-Time Spectrum of a Simple Signal

The Sound Spectograph

The short-time spectrum may be displayed with a sound spectogram -- a representation of the time-frequency-intensity coordinates of $|S(t, f)|$ [1,7]. The display is generated by playing a recorded passage of speech (typically 2.4 seconds) through a narrow band-pass filter and envelope-detector. The contour $|S(t, f_i)|$ is "burned" on Teledeltos paper for successive filter center-frequencies f_i, with relative darkness displaying intensity on a logarithmic scale. Two analyzing filter bandwidths are commonly available -- 45 Hz and 300 Hz. The choice of filter bandwidth corresponds (roughly) to the choice of the duration of the window function in equation (2-3). In the narrow-band mode, the frequency resolution is sufficient to display the voice pitch and its harmonics, but the time resolution is relatively poor. In the wide-band mode, the time resolution is sufficient to display individual glottal pulses, but the frequency resolution is relatively poor. In both modes the formant structure of the speech signal may be observed.

Narrow- and wide-band Sonagram pairs for three sentences are pictured in Figure 9. The Sonagrams in Part (a) display the spectrum of a female talker speaking the sentence "Your gift is a birthday cake." Notice the narrow vertical "line" in the spectrum produced by the burst in the /t/ in "gift" compared to the stationary nature of the /e/ in "cake." Part (b) shows the spectrum of a male talker saying "The sixth grade had a picnic." Notice the noise-like character of the fricatives /s/ and /θ/ in "sixth," and the continuity of formant transitions across word boundaries in the middle of the sentence. Part (c) shows the spectrum of a male talker saying "We were away a year ago" -- a sentence

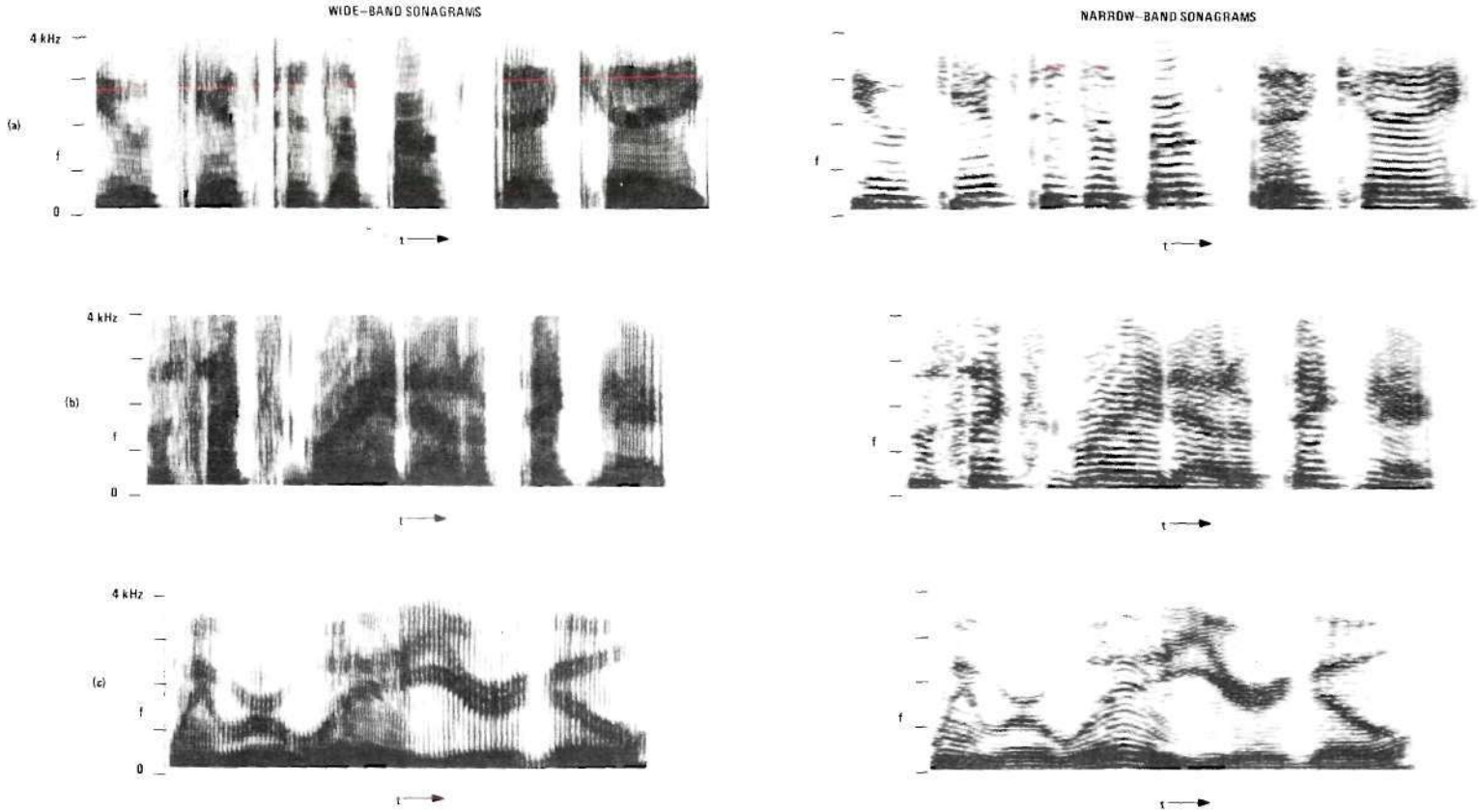


Figure 9. Wide-and Narrow-Band Sonagrams (a) Female Talker: "Your Gift is a Birthday Cake." (b) Male Talker: "The Sixth Grade had a Picnic." (c) Male Talker: "We were Away a Year Ago."

composed of non-nasal voiced phonemes. Note the slow, continuous nature of formant and pitch transitions. The vertical stirations in the wide-band Sonagrams correspond to individual glottal pulses -- the prominent peaks in each pitch period of the input signal. The narrow horizontal stripes in the narrow-band Sonagrams display the harmonic components of the voice pitch.

The Sonagrams pictured in Figure 9 were produced on a Kay Sonagraph Model 7029A Spectrum Analyzer. This machine will produce, in addition to the time-frequency-intensity display, sections of the short-time spectrum at selected time instants. A sequence of such frequency sections taken at uniform time intervals may be mounted in an array to form a three-dimensional model of the short-time spectrum surface of a speech utterance.

A spectrum model of the sentence "Noon is the sleepy time of (day)" is pictured in Figure 10. The frequency scale is 0-8 kHz, from right to left. The sections are spaced 100 ms, so that only a rough description of the short-time spectrum surface is obtained, since transitions in the surface may occur over epochs of about 10 ms. Notice the broad-band, high-frequency character of the /s/ in "sleepy" and the /t/ in "time." Continuity of formant transitions is apparent in "noon is." In producing the sections used in this model, special care was taken to get one section at the burst of the /t/ in "time," which was the most transitory feature of the spectrum surface observed on the Sonagram.

A spectrum model of the sentence "It's easy to tell the depth of a well" is pictured in Figure 11. The sections are spaced 20 ms, an

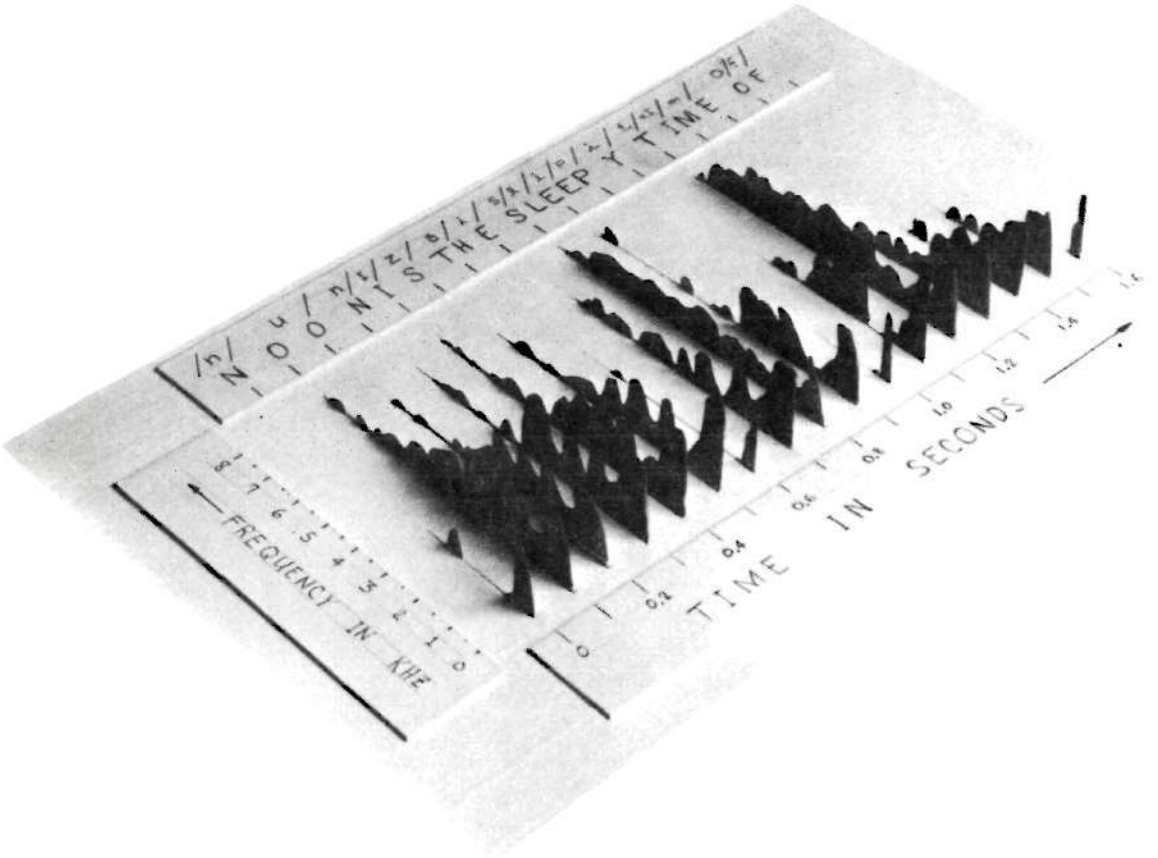


Figure 10. A Spectrum Model -- "Noon is the Sleepy Time of (Day)."

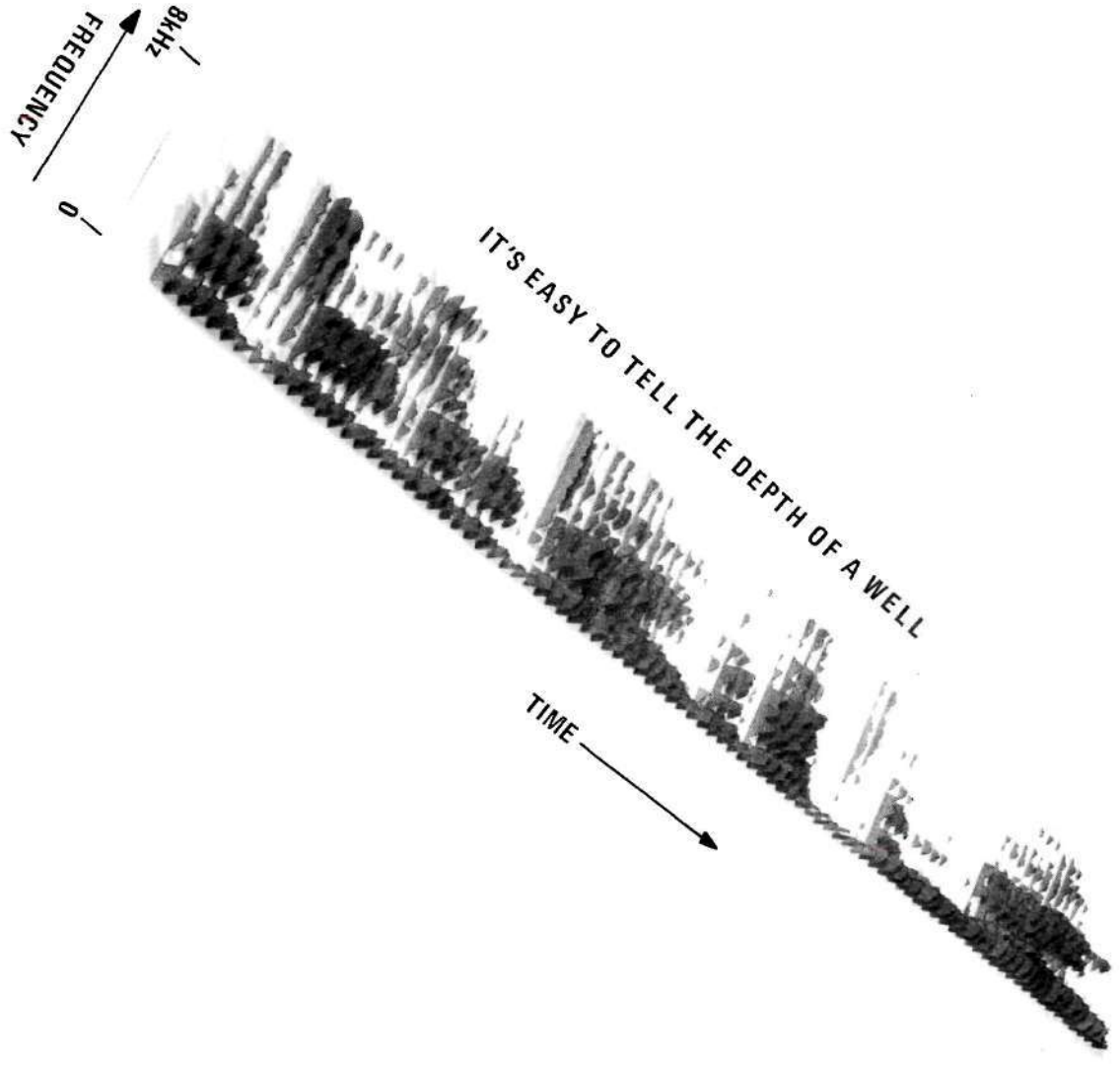


Figure 11. A Spectrum Model -- "It's Easy to Tell the Depth of a Well."

interval typical of the time "resolution" inherent to the spectrum analyzers of many vocoders. Contrast the buildup and decay of the high-frequency energy in the /s/ in "its" (which "lasts" about 80 ms) to that of the /t/ in "tell" (which appears in only one section -- about 20 ms).

The spectrum sections shown in Figures 10 and 11 display intensity on a logarithmic scale. These sections, as well as the Sonagrams pictured in this thesis, incorporate a 12 db/octave pre-emphasis above 1 kHz.

The sound spectograph has found extensive application in speech research, particularly since it provides a simple medium to communicate the subjective results of speech processing experiments. A spectographic display can "tell the story" very effectively for the vowels and voiced continuants, and is useful to a lesser degree in studying other sounds. The spectrograph offers particularly little insight into the "short" stop-consonant phonemes. Sustained, voiced sounds dominate our speech, both in intensity and duration, and are thus favored in a spectographic display, even though the "short" sounds appear to contain as much "information" as do the "long" ones. We must be cautious in basing conclusions on spectographic displays.

Digital Spectrum-Analysis

Digital spectrum-analysis is accomplished with the discrete Fourier transform (DFT) pair:

$$S(kF) = \sum_{n=0}^{N-1} s(nT)\, e^{-j2\pi nk/N} \qquad s(nT) = \frac{1}{N} \sum_{k=0}^{N-1} S(kF)\, e^{+j2\pi nk/N} \qquad (2\text{-}6)$$

where T is the sampling interval of the time function s(t), N is the number

of samples to be transformed, and $F = \frac{1}{NT}$ is the sampling interval of the spectrum [8].

To obtain the discrete short-time Fourier transform, we introduce the window function into (2-6) [7]:

$$S_r(kF) = \sum_{n=0}^{N-1} w(nT)\ s(nT + rMT)\ e^{-j2\pi nk/N} \qquad (2\text{-}7)$$

The index r corresponds to the running time variable in S(t, f). The short-time spectrum is evaluated at times t = rMT for r = 1, 2, ... The window is propagated along the time function in steps of MT seconds. The array of numbers $S_r(kF)$ given by (2-7) corresponds to the samples of S(t, f) defined in equation (2-4), to within a phase constant:

$$S_r(kF) \approx e^{+j2\pi k(rM/N-D)}\ S(rMT+D,kF); \qquad |S_r(kF)| \approx |S(rMT+D,kF)| \qquad (2\text{-}8)$$

where D is the duration of the window. The correspondence requires that the same window be used in (2-4) and (2-7) and that the window be symmetric about its midpoint and strictly time-limited, with duration $D \leqq NT$. The window function in equation (2-7) is reversed, with respect to the input time function, from that in equation (2-3). We could have avoided these differences by choosing a discrete short-time transform definition paralleling equation (2-3). But the form of equation (2-7) is more convenient and descriptive for digital processing.

Very recent papers by Oppenheim [7] and Mermelstein [9] report techniques of generating spectograms digitally by laboratory computer

using the Fast Fourier Transform (FFT) algorithm.

The FFT algorithm permits calculation of the DFT in $N \log_2 N$ complex multiplications rather than the N^2 required in direct implementation of equation (2-6). The reduction in complexity for typical DFT's of speech ranges from a factor of 18 for N = 128 to 57 for N = 512.

Two short-time spectrum sections computed digitally with the FFT are plotted in Figure 12. In Part (a) is shown a 25.6 ms segment of the waveform of the /i/ in "we," after multiplication by a Hanning window function. The corresponding amplitude spectrum plotted in Part (b) was computed with a 512-point FFT. The frequency axis runs from 0-4 kHz. Part (c) illustrates a windowed 12.8 ms segment of the /t/ in "gift is" beginning 3 ms after the initial burst. This waveform, before multiplication by the window, is plotted above in Figure 2 (a). The resulting spectrum, plotted in Part (d), displays prominent peaks near 2 and 3 kHz which in the Sonagram of Figure 9 (a) may be seen to connect with the formants of the following vowel.

The DFT in equation (6) may be motivated intuitively in several ways. Consider the sampled-data function $\bar{s}(t) = \sum_{n=-\infty}^{\infty} s(nT)\, \delta(t-nT)$ associated with the band-limited signal s(t). When the sampling frequency 1/T exceeds the Nyquist rate, the sequence $\{s(nT)\}$ is sufficient to reconstruct s(t). The Fourier transform of $\bar{s}(t)$ is

$$\bar{S}(f) = \sum_{n=-\infty}^{\infty} s(nT)\, e^{-j2\pi fnT} \qquad (2\text{-}9)$$

Suppose s(t) were essentially zero outside the interval $0 \leq t < D$, where D = NT. Then (2-9) becomes

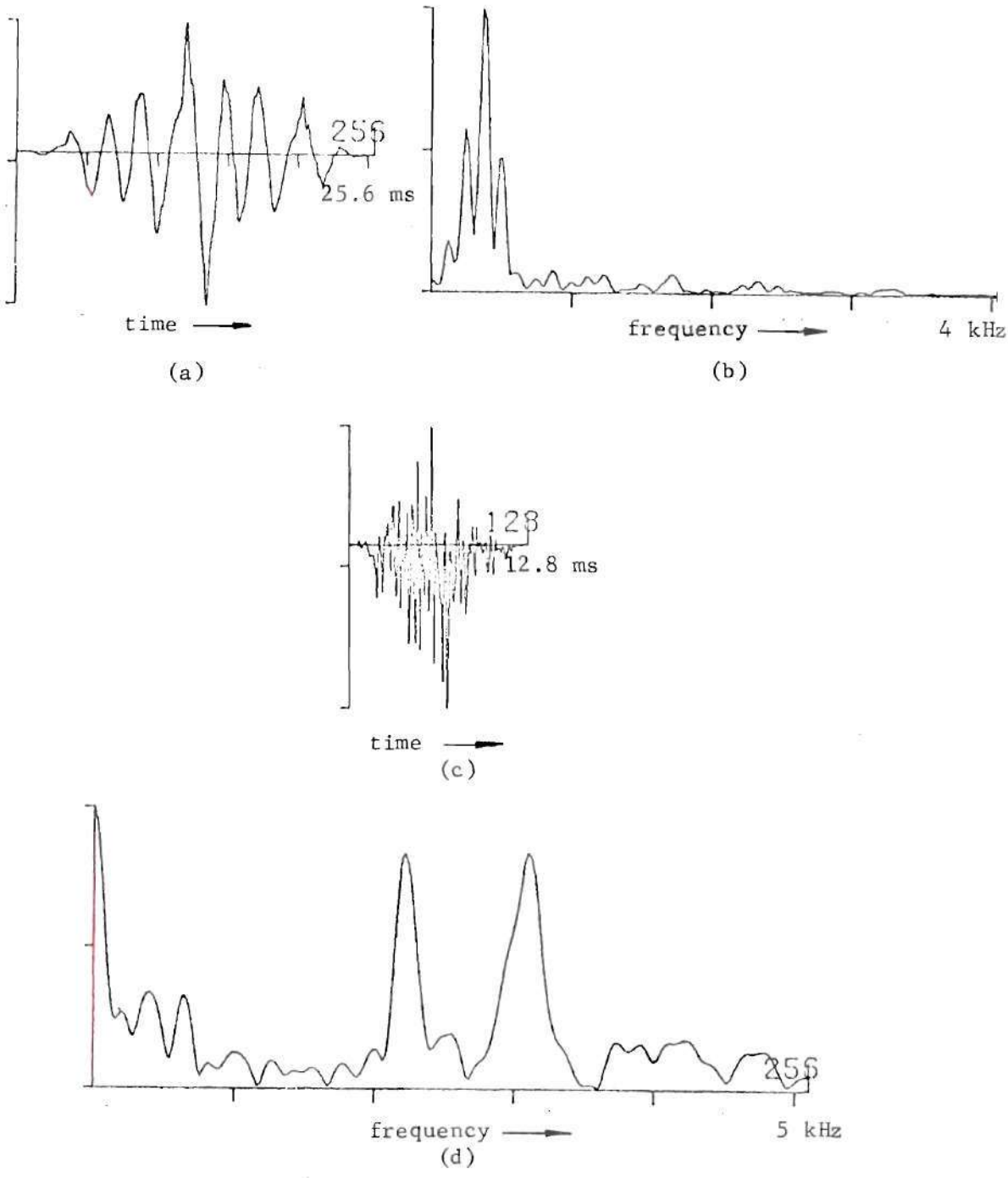


Figure 12. Spectrum Sections Obtained Digitally
(a) Windowed Waveform of the /i/ in "we"
(b) Spectrum of (a)
(c) Windowed Waveform of the /t/ in "gift is"
(d) Spectrum of (c)

$$\bar{S}(f) = \sum_{n=0}^{N-1} s(nT)\ e^{-j2\pi fnT} \tag{2-10}$$

which, when evaluated at frequencies $f = kF = k/D = k/NT$ yields the DFT:

$$\bar{S}(kF) = \sum_{n=0}^{N-1} s(nT)\ e^{-j2\pi(kF)(nT)} = \sum_{n=0}^{N-1} s(nT)\ e^{-j2\pi kn/N} = S(kF) \tag{2-11}$$

To avoid the assumption of both time and band limitedness (a mathematical impossibility), we may consider s(t) to be segmented into intervals of duration D, and each segment to be one period of a periodic signal (approximately) band limited to the same frequency as is s(t). The periodic signal is fully specified by the Fourier transform of one period, with error introduced only at the end points, where the Fourier expansion converges to the midpoint of a discontinuity. Thus, the signal s(t) may be reconstructed from the DFT's of its segments, with error introduced only every D seconds.

Another viewpoint of the DFT obtains from the z-transform. The Laplace transform of the sampled-data function $\bar{s}(t)$ is

$$\mathcal{L}\{\bar{s}(t)\} = \int_0^\infty \bar{s}(t)\ e^{-\xi t}\ dt = \sum_{n=0}^{\infty} s(nT)\ e^{-\xi nT} \tag{2-12}$$

which, after the transformation $z = e^{+\xi T}$, becomes the z-transform of the sequence $\{s(nT)\}$

$$\tilde{\bar{S}}(z) = \mathcal{L}\{\bar{s}(t)\}\Big|_{\xi = +\frac{1}{T}\ln z} = \sum_{n=0}^{\infty} s(nT)z^{-n} \tag{2-13}$$

The transformation $z = e^{\xi T}$ maps the left-half ξ-plane into the interior of the unit circle in the z-plane, and the $\xi = j2\pi f$ axis onto the $|z| = 1$ unit circle. The DFT is the sequence of N samples of the z-transform, uniformly spaced around the unit circle, just as the Fourier transform is the (double-sided) Laplace transform evaluated along the $j2\pi f$ axis (under suitable convergence conditions).

A heuristic path to the DFT is to approximate the Fourier integral by a Riemann sum:

$$S(f) = \int_{-\infty}^{\infty} s(t)\, e^{-j2\pi ft}\, dt \approx T \sum_{n=-\infty}^{\infty} s(nT)\, e^{-j2\pi fnT} \qquad (2\text{-}14)$$

For s(t) time-limited to $0 \leq t < D$, the sequence of samples at $f = k/D$ of the approximation in (2-14) yields the DFT (with proportionality constant T).

The properties of the DFT parallel those of the continuous transform. A detailed discussion is found in the book by Gold and Rader [8]. One notable difference is the DFT convolution property: the inverse DFT of the product of DFT's $U(kF) = X(kF)Y(kF)$ is the circular convolution $u(nT) = \sum_{n=0}^{N-1} X(nT)Y((n-m) \text{ modulo } N)T)$.

One interesting property of the DFT may be important in some speech processing applications. The DFT of N samples of s(t) is the set of samples of $\bar{S}(f)$ spaced $F = 1/D = 1/NT$, the maximum permissible spacing according to the frequency sampling theorem. That is, $\bar{S}(f)$ is automatically sampled at the Nyquist "rate." If a nonlinear operation follows the DFT, the resulting sequence is insufficient to interpolate the analog

of the output.

We notice that the sampling "rate" at the DFT output does not respond to an increase in the input sampling. The sampling "rate" of the DFT output may be increased by augmenting the input sequence by zeroes. For example, to double the sampling rate, we augment s(nT) with N zeroes and DFT to obtain:

$$\sum_{n=0}^{N-1} s(nT)\ e^{-j2\pi kn/2N} = S\left(k\ \frac{F}{2}\right) = \bar{S}\left(\frac{kF}{2}\right) \qquad (2\text{-}15)$$

Thus, twice as many samples of the spectrum are obtained. For even k, the samples are identical to those obtained without augmentation. For odd k, the samples are exactly the intermediate interpolated samples of the function $\bar{S}(f)$.

The discrete Fourier transform is very intimately related to the continuous Fourier transform. Since the introduction of the Fast Fourier Transform, digital spectrum analysis seems to offer great promise in the future of speech processing.

Conclusion

Short-time spectrum analysis is a very powerful tool in studying the nature of speech and in representing speech efficiently. Many of the perceptually significant features of the speech signal are apparent in the spectrum, but obscure in the waveform. Especially pleasing are the slowly changing formant peaks and pitch-harmonic contours, which we may intuitively relate to the physiology of speech production. But the short-time spectrum is not unique, since the (complex) surface produced

depends critically on the analytic form and "duration" of the window function employed. Furthermore, the method of smoothing the surface before display or coding or "feature extraction" is important (as is the subsequent interpolation in synthesis applications). Features which are obvious in one spectrum analysis may not be when a different window function is used. We return to this question in Chapter III.

The Channel Vocoder

The era of modern speech bandwidth compression research began in 1939 with the invention of the Channel Vocoder by Homer Dudley at the Bell Telephone Laboratories [3]. Dudley took the point of view of the telephone carrier engineer. He considered the sinusoidal components of the periodic pulse-train excitation of a voiced sound to be the "carriers" upon which the vocal tract imposes a "modulation" by articulation of the message. Analogously, unvoiced sounds employ a wide-band noise carrier. This point of view is essentially identical to that of the acoustic theory of speech production.

The principal strategy of the channel vocoder is to separate the influence of excitation source and vocal tract articulation. Both excitation and articulation vary at syllabic rates, so the parameters which represent them are in some sense slowly varying. The excitation is described by a voiced/unvoiced (V/UV) decision and a measure of the fundamental pitch if the sound is voiced. Articulation is described by the spectrum of the vocal tract "filter" as it changes with time. In the channel vocoder the spectrum information is obtained from the rectified and smoothed outputs of a bank of band-pass filters, spaced across the

speech band. A typical channel vocoder block diagram is shown in Figure 13 [10].

The spectrum channel signals $x_i(t)$ represent the intensity of the input speech in the vicinity of f_i -- the i^{th} filter center frequency. $x_i(t)$ is the smoothed short-time amplitude spectrum of the input $s(t)$, evaluated at f_i:

$$x_i(t) = [|S(t, f_i)|]_{LP} \tag{2-16}$$

where the symbol $[\cdot]_{LP}$ denotes the low-pass filtered version of the enclosed time function. The channel signals $\{x_i(t)\}$ are low-pass filtered to 30 Hz, so the set of 16 x_i's may be transmitted in a bandwidth of about 500 Hz. (Some "guard band" separation is needed in a frequency-division multiplexing of the channel signals.) The excitation signals require another 50 Hz, so the vocoded speech may be transmitted in a bandwidth of about 550 Hz, for a bandwidth savings of 6:1.

At the synthesizer, the inputs to a bank of band-pass filters identical to that used in the analyzer are modulated by the set of channel signals $\{x_i(t)\}$. The result is a time-varying filter whose system function varies in approximately the same fashion as the short-time spectrum of the input speech. When this filter is excited by a periodic pulse train whose pulse period is controlled by the pitch signal $f_0(t)$, a voiced sound results which has approximately the same line structure and amplitude spectrum as the original speech. Similarly, when an unvoiced excitation is indicated by the analyzer, the synthesizer filter bank is excited by noise.

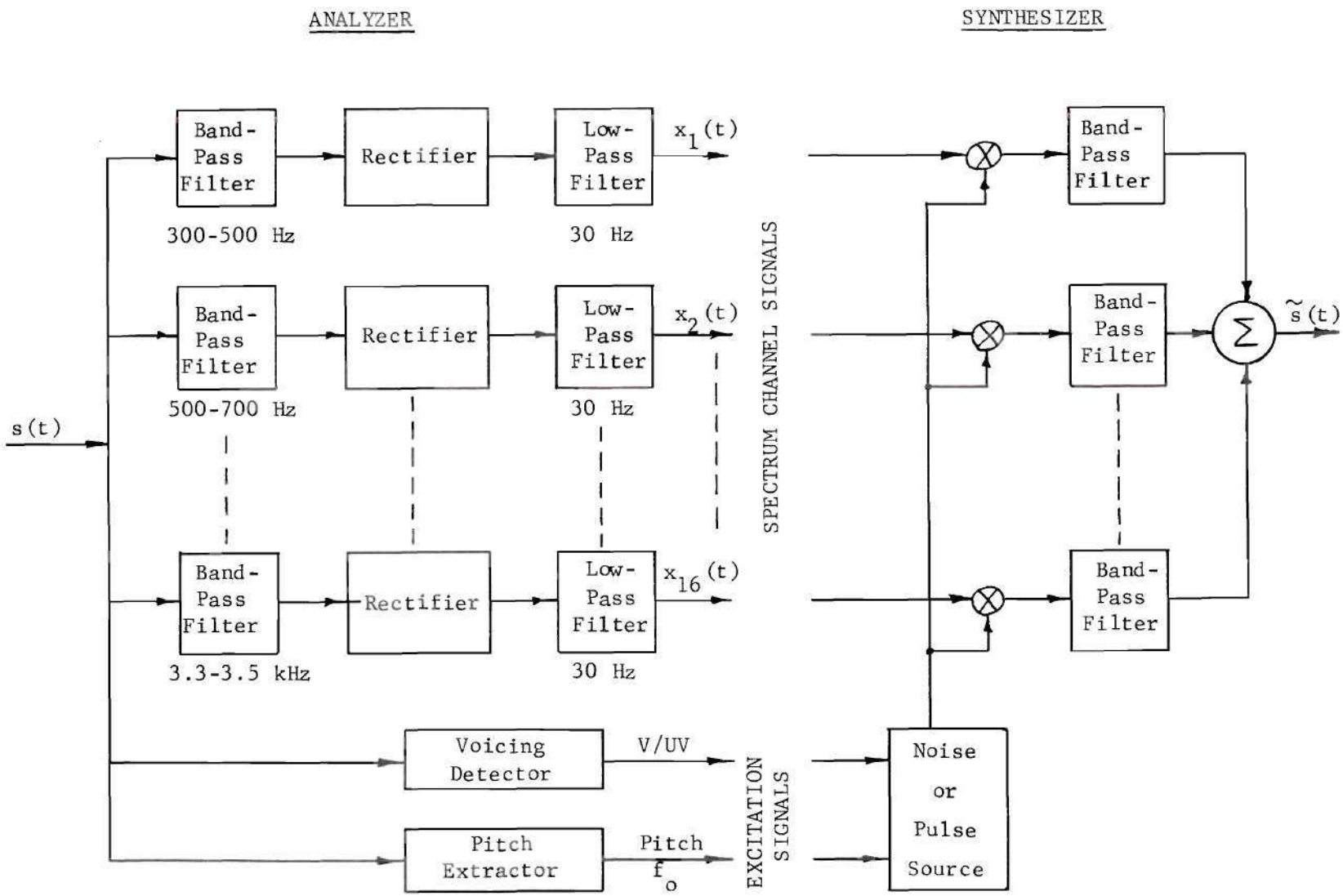


Figure 13. The Channel Vocoder

A concise way to describe the operation of the channel vocoder is that it preserves the short-time amplitude spectrum of the input speech. Note that all phase information is discarded in this scheme.

The synthetic speech produced by the channel vocoder is reported to be quite intelligible [1,10], although the quality is often somewhat machine-like and unnatural. The ability to recognize the talker may be partially lost. For these reasons, and its relatively high cost, the channel vocoder has found only a few specialized applications.

A digital channel vocoder was recently reported by Bially and Anderson which operates at a 2400 b/s rate and retains high intelligibility and speaker recognition properties [11]. A discussion of various channel vocoder realizations and design considerations is found in a paper by Gold and Rader [10].

The Formant Vocoder

A speech formant is a region of high intensity in the spectrum of a sustained sound. We may think of a formant as the manifestation of a pair of conjugate poles in the vocal tract system function. The variation of formant frequencies with time is evident in the Sonagrams of Figure 9.

The strategy of the formant vocoder is to measure the formant frequencies and amplitudes as they change with time. At the synthesizer, a parallel (or serial) bank of filters is excited in the conventional way by a synthetic voiced or unvoiced source. The filter characteristics change with time according to the received formant signals. Thus, as in the channel vocoder, the formant vocoder attempts to preserve the short-

time amplitude spectrum.

A formant vocoder block diagram is shown in Figure 14. The output of a spectrum analyzer is examined for formant peaks. The formant frequencies, F_i, and amplitudes, A_i, are measured and transmitted, along with the voicing signal(s), to the synthesizer. The center frequencies of the band-pass filters of the synthesizer are controlled by the F_i signals, and the excitation to the filters modulated by the A_i signals.

Many techniques have been devised to measure formant frequencies. The so-called "peak-picker" tracks local maxima in successive short-time amplitude spectrum sections. Schafer and Rabiner recently reported a technique of formant analysis which employs the chirp z-transform algorithm [12]. The algorithm provides an efficient way to evaluate the z-transform along a contour other than the unit circle in the z-plane. Thus, formant peaks resulting from closely spaced poles may be resolved. Considerable progress in producing synthetic speech from efficiently coded formant data was reported recently by Flanagan, et al. [2]. Weinstein and Oppenheim have obtained improvements in the data rate of the homomorphic vocoder by coding the spectrum in terms of formant parameters obtained by predictive coding [13].

The formant vocoder has inherently greater potential for bandwidth reduction than does the channel vocoder. A real-time formant vocoder whose intelligibility and quality match that of the channel vocoder has yet to be implemented [1].

Several difficulties arise in implementing a formant vocoder. During certain sounds, the unvoiced fricatives, stops, and affricates, the concept of a formant seems to break down. For such sounds, the short-

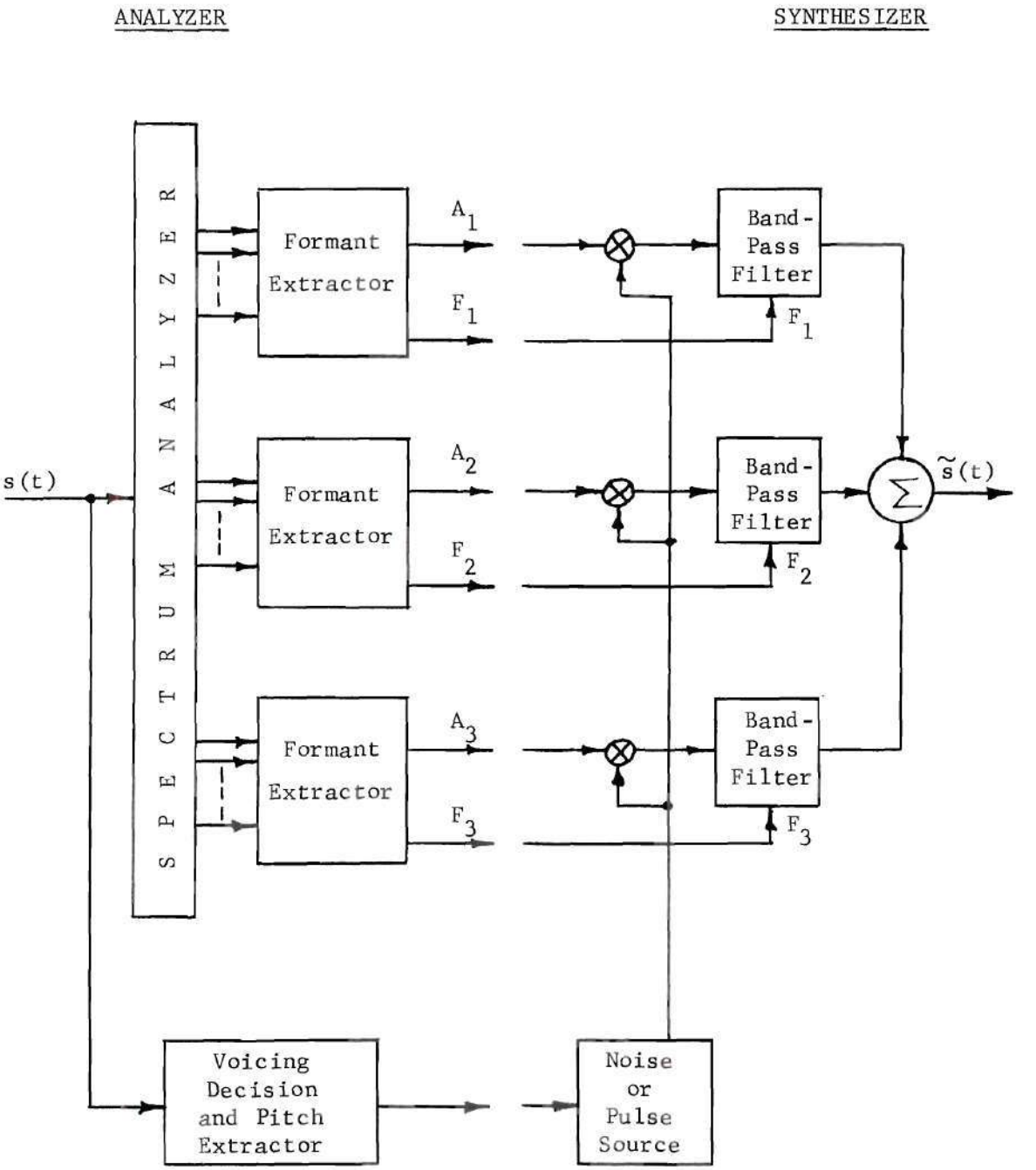


Figure 14. The Formant Vocoder

time spectrum is broad and may have no well-defined peaks. The prominent spectrum features are often quite transient in nature and thus, not easily synthesized by manipulating a few "formant filters." It is often unclear which spectrum peaks should be regarded as formants.

Although one may argue that articulatory movements are continuous, formants are often obscured by source zeroes or pauses. Thus, the formants are not continuously manifest in the spectrum, and a formant tracker is faced with both a "global" and a "local" tracking problem.

The formant coding of the short-time spectrum seems to be ideal for most sustained voiced sounds, but the stops, unvoiced fricatives, nasals, and affricates appear to require some other representation. The "generalized formant" approach implied in the recently reported work at the Bell Telephone Laboratories [2] and at MIT [13] has considerable potential.

The Homomorphic Vocoder

The homomorphic vocoder is based on an approach to nonlinear filtering due to Oppenheim [14,15,16]. The homomorphic filtering of a signal whose components are not additively combined involves the (nonlinear) transformation of the input into a domain in which the transformed components are additive. Linear filtering may then be employed. Oppenheim has obtained very high quality synthetic speech by applying the generalized linear filtering concept to the deconvolution of speech [17,18].

We motivate the strategy of the homomorphic vocoder by taking the quasi-stationary viewpoint. The spectrum of a stationary segment of speech is the product of the excitation spectrum E(f) and the vocal tract

spectrum V(f):

$$s(t) = e(t) \otimes v(t) \underset{t}{\overset{f}{\longleftrightarrow}} S(f) = E(f)\, V(f) \tag{2-17}$$

The logarithm of the amplitude spectrum $|S(f)|$ is

$$\ln |S(f)| = \ln |E(f)| + \ln |V(f)| \tag{2-18}$$

in which the influence of excitation and vocal tract are additive, suggesting linear filtering to separate the components. Taking the inverse Fourier transform of $\ln |S(f)|$ yields the so-called "cepstrum" $c(\tau)$ [19, 20]:

$$c(\tau) = \mathcal{F}^{-1} \{\ln |S(f)|\} = \int_{-\infty}^{\infty} \ln |S(f)| e^{+j2\pi ft}\, df \tag{2-19}$$

$$= \mathcal{F}^{-1} \{\ln |E(f)|\} + \mathcal{F}^{-1} \{\ln |V(f)|\}$$

During voiced segments of speech the two components of the cepstrum occupy different regions in the "quefrency" variable τ.

Since $|V(f)|$ is a "smooth" function of frequency, so is $\ln |V(f)|$, and the contribution to the cepstrum is essentially confined to the low-quefrency region on the τ axis. On the other hand, for a voiced sound $|E(f)|$ is essentially periodic in f (with peaks separated by the pitch frequency f_0), so the contribution to the cepstrum is a spike at $\tau_0 = 1/f_0$, the pitch period. A sketch of a voiced cepstrum is shown in Figure 15.

It is clear that, to the extent to which the cepstrums of E(f) and

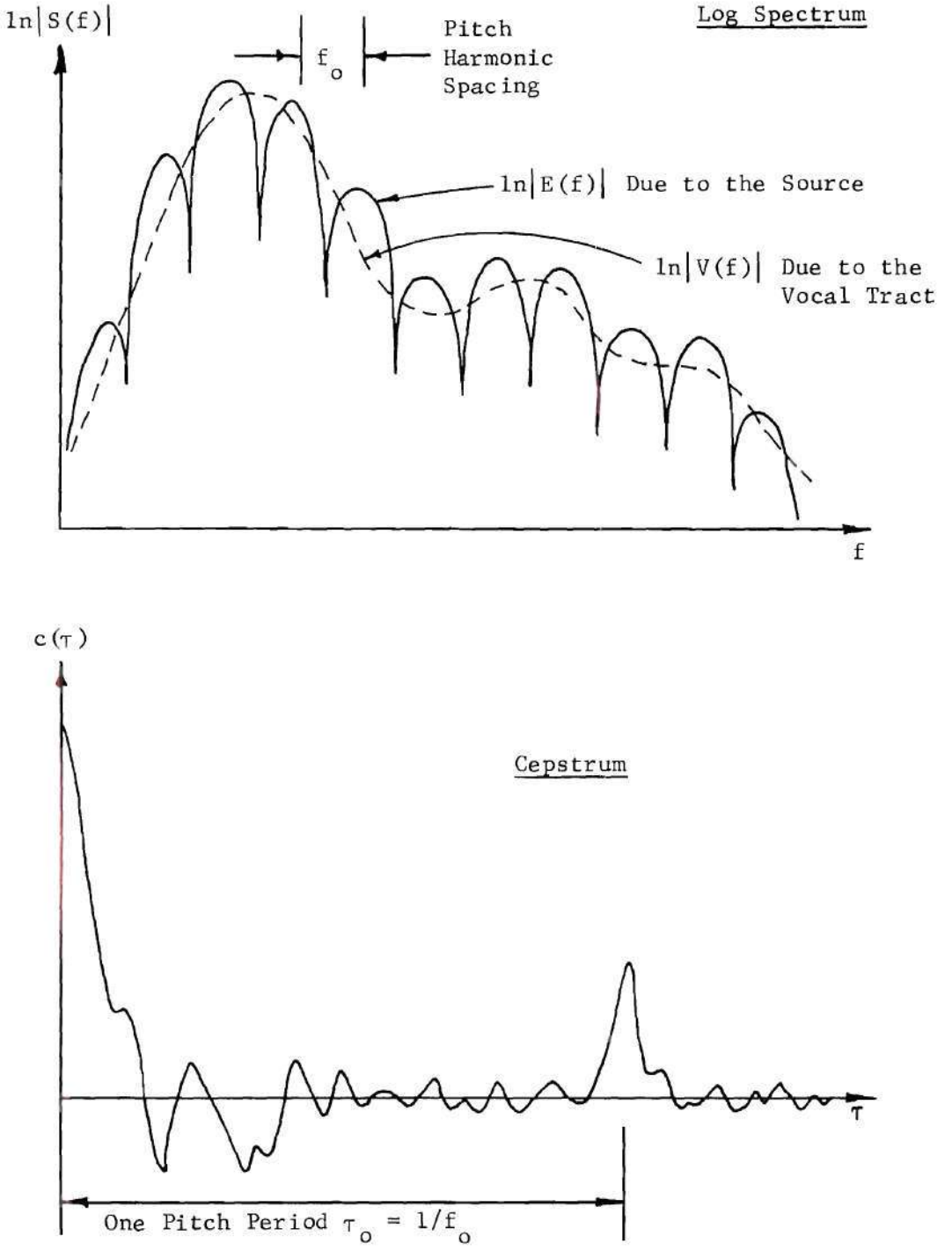


Figure 15. The Log Spectrum and Cepstrum

V(f) are disjoint in quefrency τ, the vocal tract influence may be isolated by low quefrency filtering, while the excitation may be characterized by a measure of the pitch period τ_0 in the high quefrency region. Thus, the low quefrency portion is separated for transmission as a coding of the vocal tract impulse response. Truncating the cepstrum and Fourier transforming yields the smoothed version of the log spectrum $\ln |V(f)|$. Exponentiating and inverse transforming results in a synthesized vocal tract impulse response, which may be convolved with a pitch pulse train to yield synthesized speech.

The homomorphic vocoder block diagram is shown in Figure 16. A set of plots of the waveforms encountered in the homomorphic vocoder is given in Figure 17. The operations discussed above are performed successively on short segments of the input speech. Frequency sections of the short-time spectrum are obtained (as illustrated in Figure 7 (b)) using the DFT given in equation (2-7), and the section for each successive "frame" coded by the low quefrency cepstrum for transmission.

A detailed description of the digital homomorphic vocoder operation is given by Oppenheim [18]. We only outline the description here. The input speech s(t) is sampled at $\frac{1}{T}$ = 10 kHz, and quantized, to yield the sequence s(nT), which is plotted in Figure 17 (a). (The continuous curve, produced on a Calcomp Digital Incremental Plotter, results from linear interpolation between sample points.) The 512 point (51.2 ms) segment plotted in Part (a) is from the word "your" spoken by a female, corresponding to the Sonagrams pictured in Figure 9 (a). s(nT) is multiplied by the Hanning window sequence

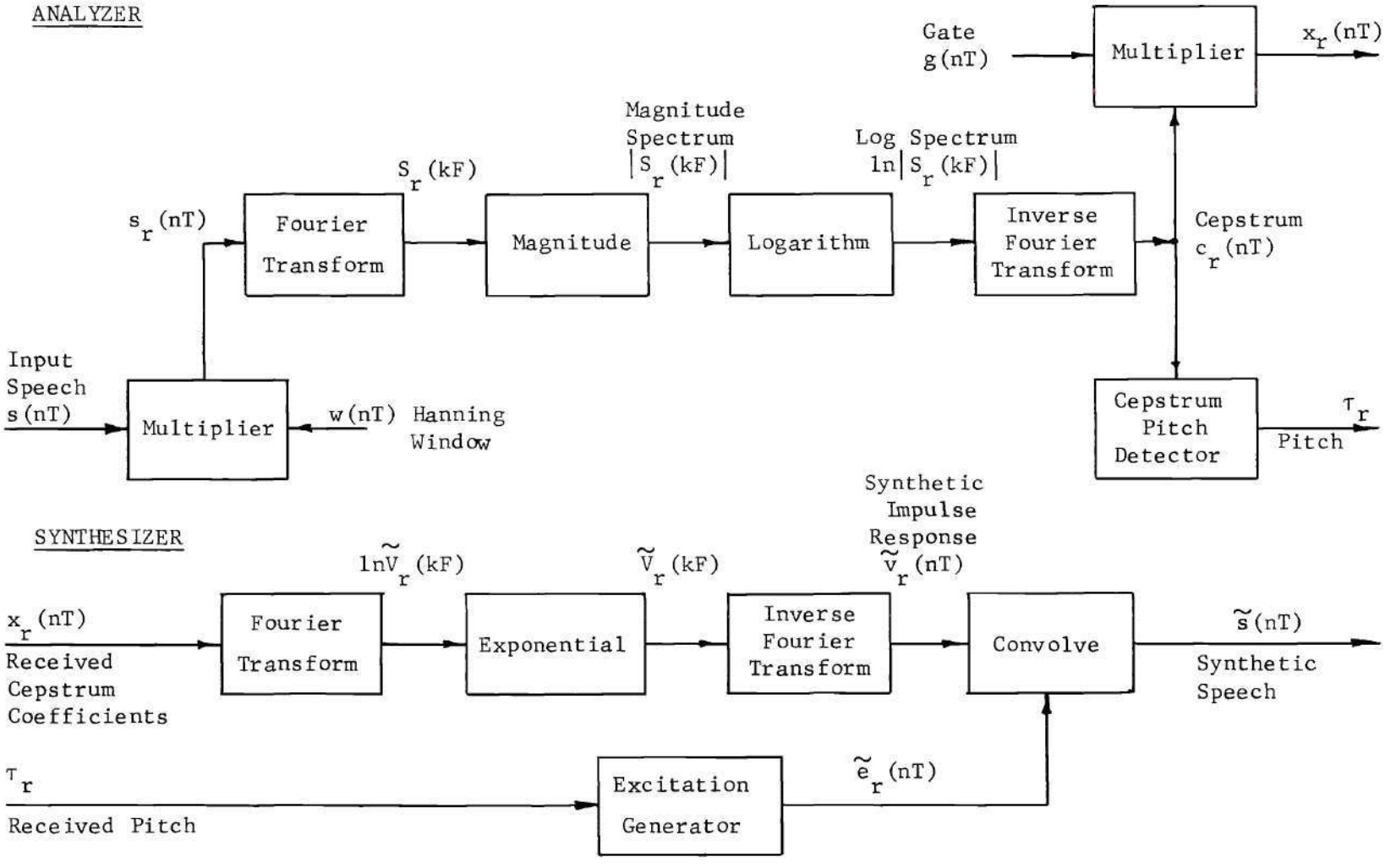


Figure 16. The Homomorphic Vocoder

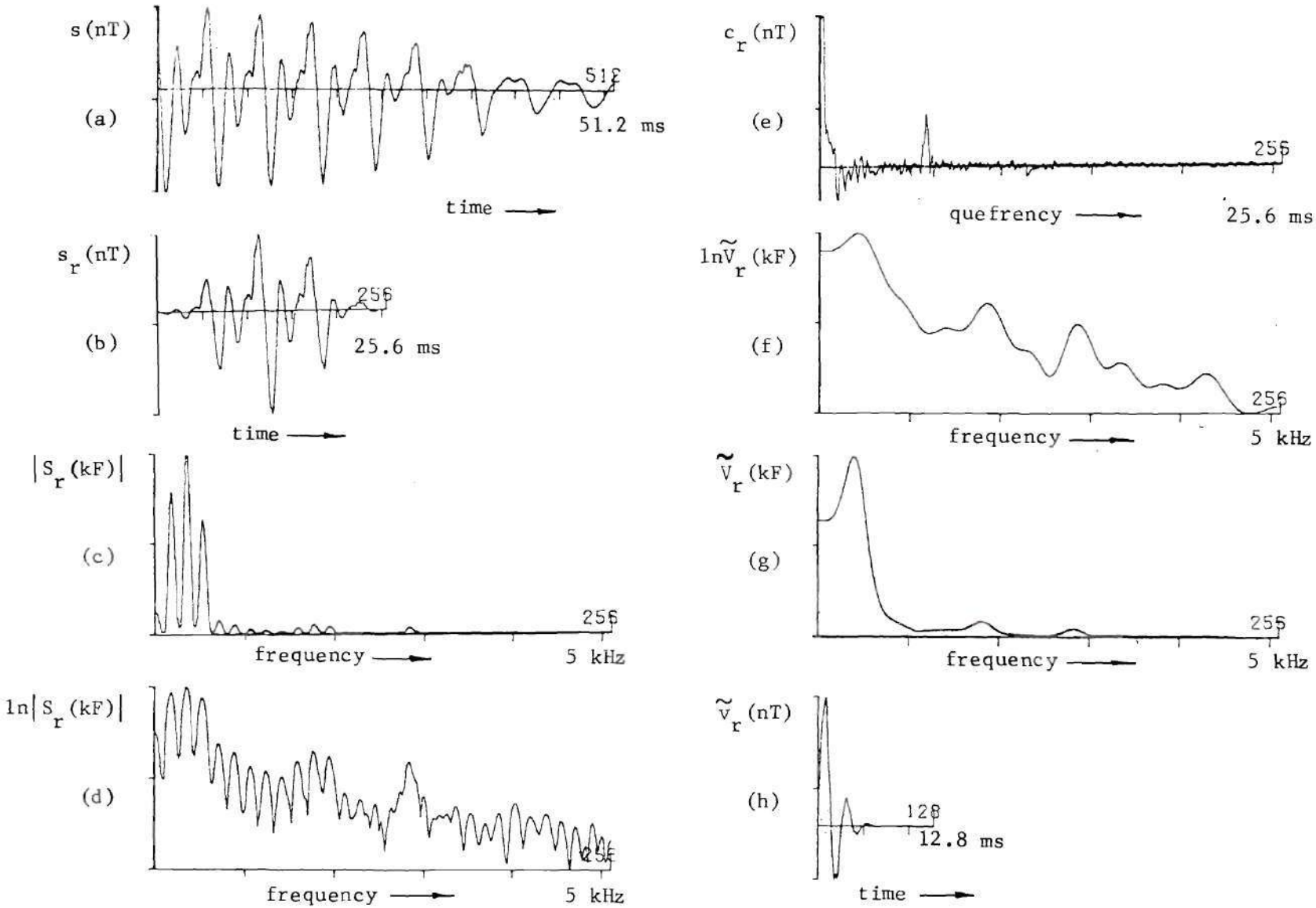


Figure 17. Typical Waveforms in the Homomorphic Vocoder

$$w(nT) = \frac{1}{2}\left(1 - \cos\frac{2\pi nT}{D}\right) \qquad 0 \leq n \leq N-1 \qquad (2\text{-}20)$$

where D = NT is the duration of the window. The windowed speech $s_r(nT) = w(nT)\, s(nT + rMT)$ for the r^{th} frame is plotted in Part (b), illustrating the use of a 25.6 ms window, N = 256. (Oppenheim used a 40 ms window.) MT, the frame propagation interval, is typically 20 ms. The DFT of $s_r(nT)$ is computed using a $2^9 = 512$ point DFT, and the magnitude taken to yield the r^{th} amplitude spectrum section $|S_r(kF)|$ plotted in Part (c). ($F = 1/NT \cong 20$ Hz; the frequency range extends to 5 kHz.) The logarithm is computed, as plotted in Part (d). Notice how the dynamic range compression of the nonlinear logarithm brings out the pitch harmonic structure at the higher frequencies. Taking the inverse DFT yields the cepstrum $c_r(nT)$, plotted in (e). The pitch period peak is clearly visible. (For convenience, the sample of the cepstrum at the origin, c(o), is not plotted, because it is of much higher amplitude than other samples. c(o) represents the "dc level" of the log spectrum.) After truncating the cepstrum to 2.6 ms, the coefficients are quantized to 6 bits ($2^6 = 64$ levels) and transmitted to the synthesizer, along with a measure of the pitch period.

At the synthesizer, the received cepstrum coefficients are Fourier transformed (by the DFT) to yield a smoothed version of the log spectrum. The smoothed log spectrum, denoted by $\ln \tilde{V}_r(kF)$, is plotted in Part (f) of Figure 17. Exponentiating yields a representation of the vocal tract system function $\tilde{V}_r(kF)$, which for convenience we call the "smoothed" spectrum. $\tilde{V}_r(kF)$, plotted in Part (g), is inverse DFT'd resulting in $\tilde{v}_r(nT)$ --

the synthesized vocal-tract impulse-response, plotted in Part (h). The excitation generator output $\tilde{e}_r(nT)$ is a unit pulse train with period τ_r for voiced sounds, or a noise-like train of pulses with random polarity for unvoiced frames. The synthesized excitation signal $\tilde{e}_r(nT)$ and impulse response $\tilde{v}_r(nT)$ are convolved to yield synthesized speech $\tilde{s}_r(nT)$. The synthesizer output is digital-to-analog converted and transduced into an acoustic signal.

The vocoder operations summarized above are performed successively on segments of the input spaced 20 ms. Since 26 coefficients of the cepstrum (each quantized to 6 bits) are transmitted for each 20 ms vocoder frame, the spectrum data rate is 7800 b/s. Quantizing the pitch signal τ_r to 6 bits requires an additional 300 b/s for excitation information. The vocoded representation of speech may be transmitted in an 8100 b/s channel, rather than the 50,000 b/s channel required to transmit the waveform representation.

The homomorphic vocoder simulated by Oppenheim [18] operates at a 7800 b/s spectrum data rate and produces very high quality synthetic speech. A real-time, homomorphic vocoder was implemented by Manley, et al. [21], on a 12 bit, fixed point processor, demonstrating that the relative complexity of the homomorphic vocoder may not inhibit its application.

The homomorphic vocoder was selected as the platform for testing the central idea in this research effort. We will return to consider the homomorphic vocoder in more detail in subsequent chapters.

Summary

Speech analysis and synthesis systems seek to achieve an efficient representation of speech by incorporating the constraints of speech production and the human hearing mechanism. The vocoders discussed in this chapter employ one constraint of speech production, namely, that vocal excitation has a broad spectrum with either quasi-harmonic (voiced) or noise-like (unvoiced) character. In addition, these vocoders incorporate the property of the hearing process that perception depends primarily upon the shape of the short-time amplitude spectrum [1].

CHAPTER III

THE TIME-FREQUENCY COMPROMISE IN THE SHORT-TIME SPECTRUM

Introduction

In this chapter we relate the uncertainty principle and scaling property of Fourier analysis and the concept of dynamic "bandwidth" and "duration" to the role of the window function in short-time spectrum analysis and the expansion of signals in both time and frequency. The notion of a resolution rectangle which describes the time and frequency properties of a spectrum analyzer is employed to examine the t-f compromise inherent in vocoder systems. We conclude that a vocoder analyzer-synthesizer with a resolution cell adapted to the nature of the signal has potential to improve the quality of vocoder systems.

The Uncertainty Principle and Window Functions

An inverse relation exists between the "duration" of a signal in time and the "bandwidth" of the corresponding spectrum. Consider $s(t) \longleftrightarrow S(f)$, a signal with unit energy, and choose the RMS definitions of duration (D_R) and bandwidth (B_R):

$$\int_{-\infty}^{\infty} |s(t)|^2 dt = \int_{-\infty}^{\infty} |S(f)|^2 df = 1, \qquad D_R^2 = \int_{-\infty}^{\infty} t^2 |s(t)|^2 dt, \tag{3-1}$$

$$B_R^2 = \int_{-\infty}^{\infty} f^2 |S(f)|^2 df$$

Applying Parseval's formula, the Schwartz inequality, and the property $S'(t) \longleftrightarrow j2\pi fS(f)$, we obtain the uncertainty principle [5]:

$$D_R B_R \geq 1/4\pi \tag{3-2}$$

Equality holds only for Gaussian signals:

$$\psi(t) = \sqrt{\frac{\alpha}{\pi}}\ e^{-\alpha t^2} \tag{3-3}$$

The value of the constant in (3-2) and the signal which minimizes the time-bandwidth product follow directly from the choice of RMS definitions of duration and bandwidth.

The uncertainty principle demonstrates the fact that we cannot limit the "occupancy" of a signal in both time and frequency below some minimum "area."

Landau and Pollak chose as a definition of duration (bandwidth) that time (frequency) interval which includes a fraction α^2 $(\beta)^2$ of the energy of the signal $s(t)$ [22,23]. This formulation leads to an uncertainty relation of the form $DB \geq \Phi(\alpha, \beta)$, where the function $\Phi(\alpha, \beta)$ may be found explicitly. The optimum function $\psi(t)$, which yields equality, has the greatest possible energy concentrated in a time-frequency cell of area DB. $\psi(t)$ is a combination of prolate spheroidal wave functions.

Another viewpoint of the inverse relation between time and frequency follows from the scaling property of Fourier transforms [5,24]:

$$s(t) \longleftrightarrow S(f) \Rightarrow s(at) \longleftrightarrow \frac{S\left(\frac{f}{a}\right)}{|a|} \tag{3-4}$$

The scaling property shows that compressing a signal $s(t)$ in time by a factor, a, expands the Fourier transform $S(f)$ by the same factor in frequency. This viewpoint is helpful when we consider the frequency resolution that may be obtained in short-time spectrum analysis using different window functions.

Short-time spectrum analysis begins with the multiplication of the input time function $s(t)$ by the window function $w(t)$. The spectrum of multiplied signals is the convolution of their spectrums. A cissoidal input $s(t) = e^{j2\pi f_1 t}$ yields a short-time amplitude spectrum

$$|S(t, f)| = |W(f - f_1)|$$

where $w(t) \longleftrightarrow W(f)$. Thus, the impulsive spectrum of a cissoid is "smeared" by convolution with the spectrum of the window. It is clear from the scaling property that a long duration window function provides the ability to resolve closely spaced frequency components of the input signal.

On the other hand, as the frequency resolution ability of the analyzer improves, the ability to resolve events in time is impaired. An impulsive input $s(t) = \delta(t - t_o)$ yields a short-time amplitude spectrum

$$|S(t, f)| = w(t_o - t)$$

so that events in time are "smeared" to the same duration as the window function.

The choice of the window function w(t) is the first crucial decision in the design of a spectrum analyzer. Window functions are selected to provide maximum concentration in time and frequency, i.e. a minimum time-bandwidth product, and for realizability in hardware. The function often selected for digital applications is the Hanning window [25]:

$$w(t) = \begin{cases} 0.5\left(1 - \cos\frac{2\pi t}{D}\right) & 0 \leq t \leq D \\ 0 & \text{otherwise} \end{cases} \tag{3-5}$$

The Hanning window combines a "smooth" time waveform with small side lobes in the spectrum. A comprehensive study of the properties of classes of window functions was recently reported by Rife and Vincent [26].

The influence of the window function on the frequency resolution obtained in spectrum analysis is illustrated in Figure 18. A segment of voiced speech was multiplied by a Hanning window of duration 25.6 ms. The resulting waveform is plotted in Part (a). The corresponding amplitude spectrum shown in Part (b) displays the pitch component (at approximately 172 Hz) and its harmonics. In Part (c) a 12.8 ms Hanning window leads to an amplitude spectrum with less frequency resolution, plotted in Part (d). The segment of speech displayed in Figure 18 is part of the utterance "your" illustrated in Figures 9(a) and 17.

Let us define the lobe bandwidth of the spectrum of a window function to be the width in frequency of the main lobe. Denote the lobe bandwidth as B_L. Similarly, the "lobe duration" of the window time-

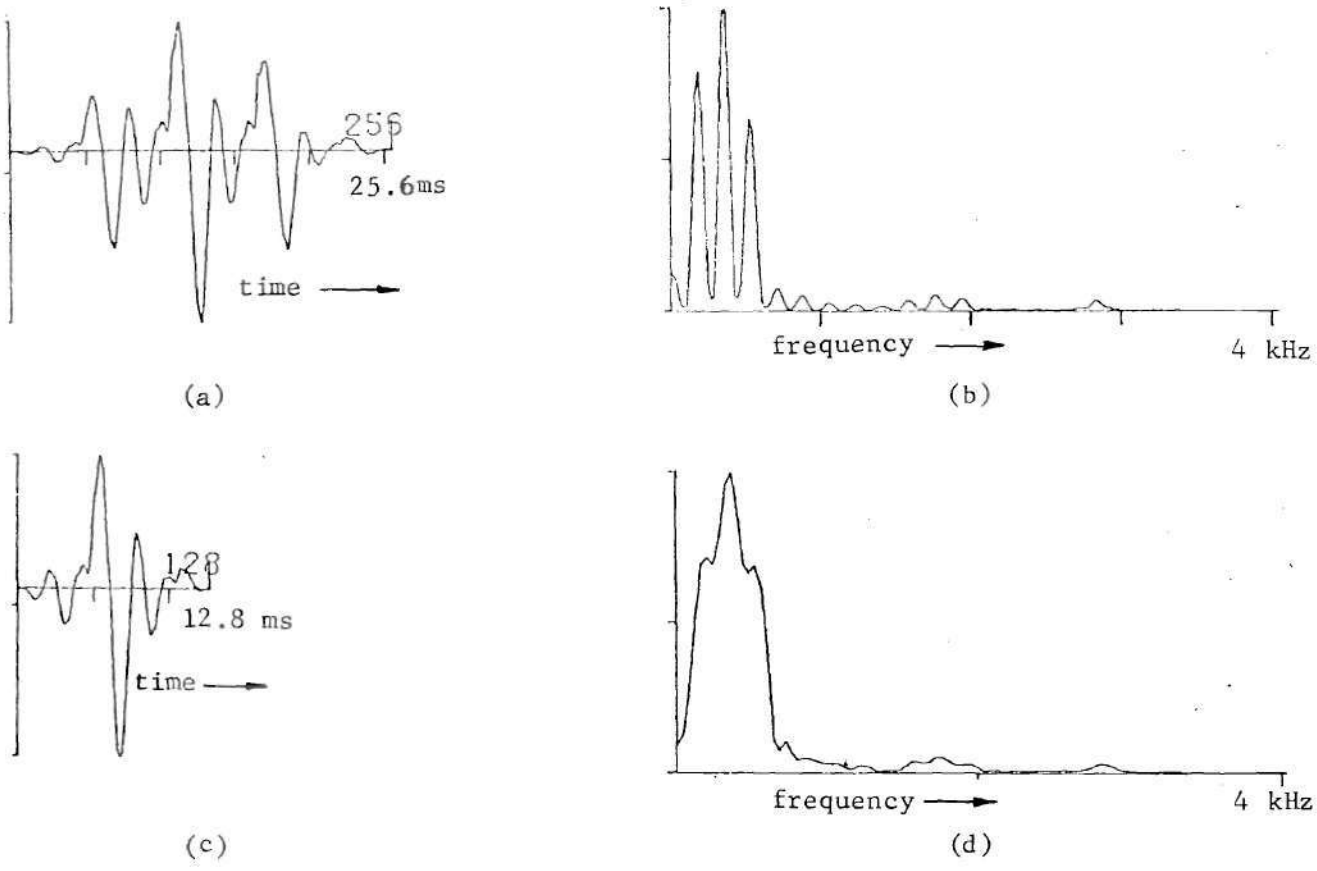


Figure 18. Effect of the Window Function on Frequency Resolution
(a) A 25.6 ms Windowed Segment of a Voiced Sound
(b) Spectrum of (a)
(c) A 12.8 ms Segment
(d) Spectrum of (c)

function is D_L. These definitions will be convenient for the particular windows discussed in this thesis, but would not be useful for windows not strictly time-limited.

The D_L = 25.6 ms Hanning window employed in Figure 18 (a) has lobe bandwidth B_L = 156 Hz. (For the Hanning window, $D_L B_L = 4.$) [25] Thus, we may consider the spectrum in (b) to have been produced by smoothing a line spectrum with a function of lobe width B_L = 156 Hz, i.e., by convolution with such a function. Thus, each "line" assumes the shape of the window's main spectral lobe. Since the lobe width (156 Hz) is less than the frequency spacing of the harmonic components, the components are well resolved. The D_L = 12.8 ms window used in Figure 18 (c), on the other hand, has lobe width B_L = 312 Hz, so the spectrum in (d) exhibits much less prominent component peaks.

We conclude that the spectrum in Figure 18 (d) has less inherent frequency resolution, and that the lobe bandwidth B_L is one useful measure of that resolution. Low frequency-resolution is not always undesirable, as all vocoders which employ the short-time spectrum representation of the vocal tract, smooth that spectrum surface to remove pitch components. Let us focus our attention on the (time and frequency) resolution properties of the short-time spectrum surface.

An impulsive input to a short-time spectrum analyzer yields at the output a ridge in the time-frequency plane of duration D_L. Similarly, a cissoidal input yields a ridge of lobe bandwidth B_L. The intersection of the responses of these "dual" time and frequency inputs is (essentially) confined to the rectangular cell of area $D_L B_L$ in the

time-frequency plane. We will refer to this cell as the resolution cell since it provides a description of the time and frequency resolution properties of a spectrum analysis performed with a particular window function.

We notice that resolution cell comparisons between spectrum analyzers which use window functions of different algebraic form are not necessarily meaningful, because of the weakness of the definition of lobe bandwidth. For example, the resolution cell of an analyzer using a rectangular window has area $D_L \times \frac{2}{D_L} = 2$, while the Hanning window leads to a cell of area $D_L B_L = 4$. But the sinx/x spectrum of the rectangular window has large side lobes which contribute considerable smearing in the short-time spectrum. The first side lobe in the spectrum of a rectangular window is 20 percent of maximum, while that of a Hanning window is only 2.5 percent of maximum.

Representation of Signals in Time and Frequency

The traditional tool used to represent a signal s(t) by a countable set of numbers is the sampling theorem [5]. If the spectrum of s(t) is zero above a frequency f_m, then s(t) can be reconstructed exactly from its samples spaced $T = 1/2f_m$:

$$s(t) \longleftrightarrow S(f) = 0 \qquad |f| \geq f_m \tag{3-6}$$

$$\Rightarrow s(t) = \sum_{n=-\infty}^{\infty} s(nT) \frac{\sin[2\pi f_m (t - nT)]}{2\pi f_m (t - nT)}$$

The "dual" of the sampling-theorem expansion is the Fourier series expansion [5]. If s(t) is time limited to a duration D, then s(t) can be reconstructed exactly from samples of its spectrum spaced F = 1/D:

$$s(t) \longleftrightarrow S(f) \qquad s(t) = 0 \qquad |t| \geq \frac{D}{2} \tag{3-7}$$

$$\Rightarrow s(t) = \begin{cases} \frac{1}{D} \sum_{k=-\infty}^{\infty} S(kF)\ e^{+j2\pi kFt} & |t| < \frac{D}{2} \\ 0 & \text{otherwise} \end{cases}$$

These "dual" expansions are only approximated in communication system practice since only finite-length sequences may be processed.

The important concept of dimensionality of a signal s(t) is illustrated by both expansions. From (3-6) we observe that if s(t) were both band limited to f_m and time limited to a duration D, then s(t) could be represented by $2f_mD + 1$ real numbers. (We assume the signal s(t) to be real.) The same observation follows from (3-7), since for real s(t), $S(f) = S^*(-f)$. In other words, only $2f_mD + 1$ independent signals $\phi_i(t)$ may "occupy" a (single-sided) bandwidth f_m and duration D. The "mathematical truth" in this "engineering intuition" was examined in detail by Landau and Pollak [23].

The Gabor Expansion

An early attempt to represent signals in both time and frequency was reported by Gabor in 1946 [27]. His classic paper "Theory of Communication" described an inquiry into the essence of the "information"

conveyed by communication channels.

Gabor introduced the analytic signal formulation [28]:

$$z(t) = s(t) + j\,\check{s}(t) \qquad \check{s}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}\,d\tau \qquad (3\text{-}8)$$

$z(t)$ is the analytic signal associated with $s(t)$. $\check{s}(t)$ is the signal in quadrature with $s(t)$, computed with the Hilbert transform. Equivalently, $\check{s}(t) = s(t) \otimes \frac{1}{\pi t}$, and since $\frac{1}{\pi t} \longleftrightarrow -j \operatorname{sgn} f$, the spectrum of $z(t)$ is a one-sided version of $S(f)$:

$$Z(f) = 2S(f)\,U_{-1}(f) \qquad (3\text{-}9)$$

where $U_{-1}(f)$ is the unit step function. $S(f)$ is completely specified by $Z(f)$ because $s(t)$ is real, with symmetry in its spectrum $S(-f) = S^*(f)$. Similarly $s(t)$ may be recovered from $z(t)$ by taking the real part. Analytic signals are especially useful in modulation problems.

The utility of analytic signals in studying time and frequency is that the single-sided spectrum $Z(f)$ permits a reasonable definition of RMS bandwidth as the second central moment of a band-pass spectrum. The discussion of the uncertainty principle which led to equation (3-2) implicitly assumed that the time function $s(t)$ was centered at the origin, and that its spectrum $S(f)$ was low-pass.

Using the second central moment definitions of RMS bandwidth and duration, Gabor obtained an uncertainty relation

$$D\,B \geq \frac{1}{2} \qquad (3\text{-}10)$$

where the lower bound of one-half resulted from a scale factor $\sqrt{2\pi}$ employed in the definitions of D and B.

The signal which minimizes the time-bandwidth product (3-10) is the modulated Gaussian pulse:

$$\psi(t) = \exp\left[-\alpha^2(t - t_o)^2\right] \exp\left[j(2\pi f_o t + \Phi)\right] \tag{3-11}$$

where α determines the "sharpness" of the pulse in time, t_o the epoch of the peak, and f_o and Φ the frequency and phase of the modulating oscillation. A Gaussian time function has a Gaussian spectrum. Thus, we picture the pulse $\psi(t)$ as producing a "hill" in the short-time amplitude spectrum t-f plane. The "hill" has Gaussian vertical cross sections, and is centered at (t_o, f_o).

Since the "elementary signal" $\psi(t)$ is "optimum" in the sense that it "occupies" minimum area in the t-f plane, Gabor chose $\psi(t)$ as the building block for expanding arbitrary signals. ($\psi(t)$ is optimum only with respect to the particular RMS definitions of duration and bandwidth used by Gabor.) Gabor proposed an expansion of the analytic signal z(t) in the time and frequency translated versions of $\psi(t)$:

$$z(t) \cong \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{nk} \exp\left[\frac{-\pi(t - nD)^2}{2D^2}\right] \exp\left[\frac{j2\pi kt}{D}\right] \tag{3-12}$$

The coefficients c_{nk} are, roughly speaking, the complex samples of z(t) in time and frequency. Each c_{nk} represents z(t) in a rectangular region of the t-f plane of dimensions $D \times \frac{1}{D}$, illustrated in Figure 19.

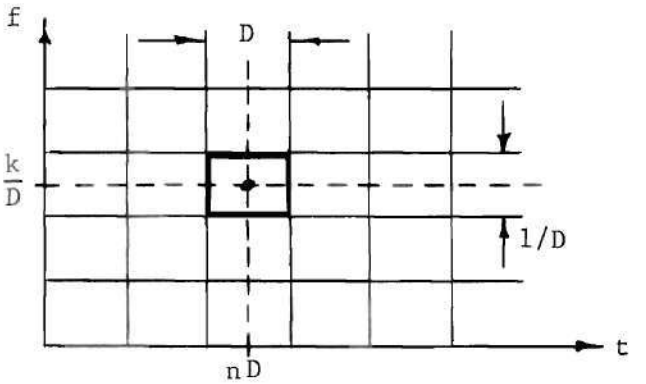


Figure 19. Unit Cells in the Time-Frequency Plane

The elementary signals $\psi(t)$ are not orthogonal, but overlap their neighbors in both time and frequency. Thus, the coefficients c_{nk} must be obtained by numerical methods. For this reason, the Gabor expansion is useful mainly as a conceptual tool.

An exact continuous counterpart to the Gabor expansion was reported by Helstrom in 1966 [29]. The "elementary signal" kernel corresponding to (3-11) is

$$\psi(t;\ \tau,\ f) = (2\pi D^2)^{-1/4} \exp\left[\frac{-(t-\tau)^2}{4D^2}\right] \exp\left[j2\pi f\left(t - \frac{\tau}{2}\right)\right] \qquad (3\text{-}13)$$

The exact version of (3-12) is

$$z(t) = \int_{-\infty}^{\infty}\int g(\tau,\ f)\ \psi(t;\ \tau,\ f)\ d\tau df \qquad (3\text{-}14)$$

where the expansion function $g(\tau,\ f)$ is given by

$$g(\tau, f) = \int_{-\infty}^{\infty} \psi^*(t; \tau, f)\, z(t)\, dt \qquad (3\text{-}15)$$

Replacing the double integral in (3-14) by a double summation leads to (3-12). The self-convolution of $g(\tau, f)$ is proportional to the ambiguity function of $z(t)$.

The Lerner Expansion

Gabor's conceptual time-frequency expansion was generalized by Lerner to allow for the use of a less restricted elementary signal [30]. Lerner proposed an expansion in unit functions $v_{mn}(t)$ which are the time and frequency translates of some "convenient" signal $v(t)$,

$$v_{mn}(t) = v(t - nD) \exp\left(j \frac{2\pi mt}{D}\right) \qquad (3\text{-}16)$$

An integer step in the index n shifts the epoch of $v(t)$ one unit -- D in time, while an integer step in index m shifts the spectrum of $v(t)$ by one unit -- $1/D$ in frequency. A convenient $v(t)$ is one which has a small time-bandwidth product. Further, we may desire $v(t)$ to be the impulse response of a realizable filter. The expansion for any "well-behaved" signal $s(t)$ is

$$s(t) \cong \sum_m \sum_n a_{mn} v_{mn}(t) \qquad (3\text{-}17)$$

A proof is given by Lerner [30].

We notice that, if the translation interval D is made small enough, $1/D$ exceeds the highest frequency component in the spectrum of $s(t)$, and the expansion becomes a sampled time representation. Similarly,

increasing D until it exceeds the duration of the signal s(t) results in a Fourier expansion.

The Gabor-Lerner expansions and the Helstrom transform were generalized by Montgomery and Reed [31], who generalized Helstrom's expansion as Lerner did Gabor's. That is, the Gaussian envelope of the kernel is replaced by any "convenient" square-integrable function.

The Short-Time Spectrum

We may think of the short-time spectrum as a coding of the coefficients of a Gabor-Lerner expansion. The window function w(t) in the short-time spectrum corresponds to the "elementary signal" $\psi(t)$ or the convenient expansion signal v(t) in the Gabor and Lerner expansions. The expansions seek to represent a waveform exactly and offer no reduction in the dimensionality required in the coding. Conversely, the short-time amplitude spectrum is employed in the vocoder as a coding of the vocal tract articulation.

The dimensions of the cell in time and frequency ($D \times 1/D$) which characterize the Gabor-Lerner sampling pattern suggest that the sampling pattern of the short-time spectrum should be matched to the dynamics of the articulation to be represented. Perhaps the duration of the window function should correspond to the relative "duration" of the articulatory event to be approximated. That is, perhaps we should match the shape of the time-frequency cell represented by a sample of the short-time spectrum to the nature of the segment of speech being processed.

Short-Time "Bandwidth" and "Duration"

In a recent paper entitled "Signal Energy Distribution in Time and

Frequency" Rihaczek examined the relation of the Fourier representation of signals to the signal structure in time and frequency [32]. Rihaczek derived a function $\varepsilon(t, f)$ -- the complex energy density function -- which appears to be useful in studying the distribution of energy in the t-f plane:

$$\varepsilon(t, f) = z(t)\, Z^{*}(f)\, e^{-j2\pi ft} \tag{3-18}$$

where $z(t)$ is the analytic signal associated with the real signal $s(t)$, i.e. $z(t) = s(t) + j\check{s}(t)$.

Suppose the signal $s(t)$ has energy E, i.e.

$$\int_{-\infty}^{\infty} s^2(t)\, dt = \frac{1}{2} \int_{-\infty}^{\infty} |z(t)|^2 dt = E \tag{3-19}$$

The integral of $\varepsilon(t, f)$ over the entire time-frequency plane yields the energy:

$$\frac{1}{2} \iint_{-\infty}^{\infty} \varepsilon(t, f)\, dtdf = E \tag{3-20}$$

Integrating $\varepsilon(t, f)$ over time produces the energy density spectrum,

$$\int_{-\infty}^{\infty} \varepsilon(t, f)\, dt = |Z(f)|^2 \tag{3-21}$$

whereas the integral over frequency yields the energy density waveform

$$\int_{-\infty}^{\infty} \varepsilon(t, f)\, df = |z(t)|^2 \tag{3-22}$$

The Fourier transform of the autocorrelation function of a signal is its energy density spectrum $|Z(f)|^2$, and the Fourier transform of the autocorrelation function of a signal spectrum is the energy density waveform $|z(t)|^2$. The complex energy density function matches these "dual" results. That is, the double Fourier transform of the combined autocorrelation function in time and frequency is $\varepsilon(t, f)$. The two-dimensional correlation function is essentially equivalent to the radar ambiguity function.

$\varepsilon(t, f)$ is the complex energy density, so its value at a particular point (t_o, f_o) is not itself a measure of the energy distribution, since t_o may fall at a zero crossing of $z(t)$, or a positive energy density at one point may be offset by a nearby negative region, so that little energy may actually be concentrated where $|\varepsilon(t, f)|$ is large. Rather, the energy distribution is described by the integral of $\varepsilon(t, f)$ over a cell in time and frequency, where the dimensions of the cell are chosen large enough to prevent significant interaction of $\varepsilon(t, f)$ over adjacent cells. This cell relates to the cells in the t-f plane "occupied" by an elementary signal of the Gabor-Lerner expansion.

Writing signal and spectrum in terms of amplitude and phase, $z(t) = |z(t)|e^{j\Phi(t)}$, $Z(f) = |Z(f)|\ e^{j\theta(f)}$, the energy within time interval T and frequency interval B is

$$E_{T,B} = \frac{1}{2}\iint_{T,B} |z(t)|\ |Z(f)|\ e^{j[\Phi(t) - \theta(f) - 2\pi ft]}\ dtdf \qquad (3\text{-}23)$$

The significant contributions to the integral occur near the points of stationary phase:

$$f = \left(\frac{1}{2\pi}\right) \frac{d\Phi(t)}{dt} = f_i(t) \tag{3-24}$$

$$t = -\left(\frac{1}{2\pi}\right) \frac{d\theta(f)}{df} = \tau_g(f)$$

where $f_i(t)$ is the "instantaneous frequency" and $\tau_g(f)$ the "group delay."

The dimensions of the cell within which energy is concentrated are obtained by approximating the time (frequency) interval within which the phase $\Phi(t)$ ($\theta(f)$) deviates by $\pi/4$ from linearity, using the quadratic term in the Taylor expansion of $\Phi(t)$ ($\theta(f)$). The results are

$$T_r = \left[\frac{2\pi}{\ddot{\Phi}(t)}\right]^{1/2} = \left[\frac{df_i(t)}{dt}\right]^{-1/2} \tag{3-25}$$

$$B_d = \left[\frac{2\pi}{\theta''(f)}\right]^{1/2} = \left[\frac{d\tau_g(f)}{df}\right]^{-1/2}$$

T_r is the time interval over which the signal phase may be considered linear, or the instantaneous frequency constant, so it may be called the short-time duration. (Rihaczek called T_r the relaxation time.) Analogously, B_d is called the dynamic signal bandwidth. To determine the size of the cell $T_r \times B_d$, Rihaczek obtained the result

$$B_d = \left[\frac{\ddot{\Phi}(t)}{2\pi}\right]^{1/2} = \frac{1}{T_r} \qquad \text{at } t = \tau_g \tag{3-26}$$

That is, the cell area is unity, $T_r \times B_d = 1$, a result reminiscent of the uncertainty principle and Gabor's subdivision of the time-frequency plane with elementary signals.

Rihaczek's results suggest that the vocoder approximation of the short-time amplitude spectrum of speech might be improved by matching the t-f resolution cell of the analyzer and the t-f sampling pattern to the short-time duration and dynamic bandwidth of the "vocal tract signal" to be represented.

The Time-Frequency Compromise in Vocoder Design

Let us focus our attention on the vocal tract spectrum information employed in vocoder analysis-synthesis. Each of the vocoders discussed in Chapter II has the property that a fixed time-frequency (t-f) compromise is inherent in its design.

In the channel vocoder sixteen 200 Hz bandwidth spectrum channels cover the range 300-3500 Hz. The band-pass filters give 200 Hz resolution in frequency, and their impulse response has duration $D \approx 5$ ms. The smoothing low-pass filter has duration $D \approx 15$ ms. The combined memory of the two filters is about 20 ms. In a typical digital channel vocoder, the spectrum channels are sampled every 20 ms. Such a vocoder encodes the vocal tract information as the samples in the t-f plane of the smoothed short-time amplitude spectrum, one sample for each rectangular cell of dimensions 20 ms $\times$ 200 Hz.

Let us define the resolution rectangle as a cell in the t-f plane within which a spectrum analyzer is unable to separate signals. (The resolution cell we considered above described the limitation on time and frequency resolution inherent in the complex spectrum due to the window function.) The resolution rectangle describes the resolution inherent in a representation of the short-time amplitude spectrum (surface) including the effects of the window function and the subsequent smoothing. The

uncertainty relation for Fourier transforms and signals ("Duration" × "Bandwidth" ≧ 1/2) suggests that the smallest possible resolution rectangle has area 1/2, and that the duration and bandwidth dimensions of the rectangle are inverse quantities. In the vocoder the area of the resolution rectangle is determined by the allocated bandwidth or data rate for vocal tract information. In the channel vocoder considered above, any t-f sampling pattern which yields a resolution cell area $D \times B = 4$ results in the same sample rate (800 samples/sec).

In the homomorphic vocoder the initial t-f compromise is the choice of the window function, since its shape and duration determine the resolution cell for the computed spectrum. The input speech is weighted by a 40 ms duration Hanning Window followed by a discrete Fourier transform operation yielding a high resolution amplitude spectrum with resolution cell DB = 40 ms × 100 Hz = 4. Time gating (i.e., low quefrency filtering) the cepstrum to 2.6 ms corresponds to a linear smoothing of the log-amplitude spectrum, reducing the frequency resolution to about 190 Hz. The 40 ms window is advanced in 20 ms steps along the input time function, and a new cepstrum frame transmitted every 20 ms. Due to this overlap, successive cepstrums are in some sense averaged over time, but since the "tails" of the Hanning window are small, we may consider the homomorphic vocoder to achieve approximately 20 ms time resolution. (This is a rough approximation, because a "short" time event in the waveform would fall into the window for two successive frames, so in synthesis the "short" event would be "smeared" to a duration of 40 ms.)

We conclude that the homomorphic vocoder employs a fixed t-f compromise with a resolution rectangle of approximately D × B = 20 ms ×

190 Hz = 3.8. Three factors determine the resolution rectangle:

1. the duration (and shape) of the input window function,
2. the amount of overlap between successive input frames, and
3. the number of cepstrum parameters transmitted per frame.

What is the optimum compromise in t-f resolution? The analysis-synthesis strategy of the vocoder is based on the stationary model of speech production. It seems clear that the stationary model is only valid to the extent that the sound viewed through the analysis window is stationary. A study of Sonagrams of conversational speech indicates that the spectrum features of some "short" sounds (e.g. the stop consonant /t/) "last" for only 15 or 20 msec while vowel sounds appear stationary for 60 msec or longer. The apparent "duration" of these features in a short-time spectrum analysis includes the memory of the analyzing filter, that is, the time resolution. Examination of the Calcomp plots of conversational speech reveals that the burst of the stop consonants normally decays within 10 ms.

There is evidence that the ear makes a mechanical short-time frequency analysis at an early stage of processing. Flanagan has advanced a filter model for this observed effect [1]. Thus, in the human hearing process, the ear "filter" may be able to trade between time and frequency resolution in accordance with the uncertainty relation. Some experimental evidence supports this notion.

Malme measured the ability of a human observer to discern a peak or valley in an otherwise flat noise spectrum [33]. Spectral peaks with Q's less than about 5 and spectral valleys with Q's less than about 8

were found to be not perceptible. This result suggests that the hearing process is rather insensitive to the detailed shape of the spectrum of noise-like sounds.

Flanagan measured the difference limen (i.e. the just-noticeable difference) of a vowel formant frequency to be on the order of three to five percent of the formant frequency [1]. The difference limen (DL) for first formant amplitude is roughly 1.5 db, while the DL for first formant bandwidth is about 20 percent. We may conclude that the hearing process is rather sensitive to the detailed shape of the spectrum of vowel sounds.

After a survey of experimental results related to the perception of speech, Flanagan writes "The data in the preceding discussions suggest that speech perception is an adaptive process. It is a process in which the detection procedure probably is tailored to fit the signal and the listening task." [1]

Lecours and Sparkes compared the performance of two spectrum analyzers with different resolution rectangles (10 ms $\times$ 100 Hz and 5 ms $\times$ 200 Hz) in the pattern recognition of stop consonants and vowels. In general, the 100 Hz mode was superior for vowel sounds while the 200 Hz mode was superior for the stop consonants [34].

Pyron and Williamson studied the properties of the stop consonants by electronic blanking of short segments to determine the distribution in time of recognition clues to perception [35]. They found that the recognition clues for some stop consonants are heavily concentrated in the initial 10-15 ms of the waveform. We conclude that the vocal tract impulse response to be used for synthesis of such a sound should be obtained from a spectrum analysis with window function duration no longer

than about 15 ms.

Voiers evaluated eight state-of-the-art, 2400 b/s, digital vocoders in 1968 using the Diagnostic Rhyme Test [36]. All of the systems tested performed well with respect to four of six attributes of consonant phonemes, but none permitted adequate discrimination of the attributes sustention (fricative versus stop consonant) and graveness. The most significant deficiency was in sustention. Not only was the discriminability score lowest for this attribute, but a drastic bias was indicated in favor of the sustained (fricative) state of the attribute. Voiers concludes: "It is hypothesized that these deficiencies stem primarily from inadequate spectrum sampling rate (i.e., frame rate) rather than inadequate 'spectrum resolution'."

The stationary model for speech production is central to the vocoder concept. Vocoder analysis-synthesis can only reproduce the short-time spectrum features of those sounds which are, indeed, stationary (or quasi-periodic) over some interval, and such reproduction requires that the time resolution capability of the processor approximate the duration of the quasi-stationary intervals. It is clear that the spectrum derived from a window which overlaps significant portions of adjacent phonemes of different character (e.g. a /t/ followed by a vowel) will either be dominated by one phoneme (the more intense vowel) or represent a corrupted average of dissimilar spectrums. A 40 ms window "propagating past" a /t/ will smear the 10 ms burst to 40 ms in the short-time spectrum. The stationary model, and thus the ability of a vocoder to reproduce the spectrum of the input speech, is only reasonable to the extent that the successive sounds "viewed" through the window are stationary.

We reach the conclusion that there is no "optimum" compromise in time-frequency resolution. Instead, the "filter" nature of the hearing process, the extremes in the articulatory dynamics of speech production, the desire for validity of the stationary model, and the concept of a t-f cell "matched" to the signal suggest that the shape of the resolution rectangle in vocoder spectrum analysis should be adapted to the signal. "Short" sounds, whose detailed spectral shape are relatively unimportant in perception, should be analyzed with a "short" window (good time resolution), while the sustained vowel sounds (whose detailed spectral shape is important) should be analyzed with a "long" window (good frequency resolution). This conclusion is summarized in Figure 20.

Thus, it appears that an improvement in vocoded speech quality and intelligibility may be achieved using a vocoder analysis-synthesis strategy which adapts the resolution rectangle to the nature of the segment of speech being analyzed. Alternately, the spectrum sampling rate might be reduced.

Conclusion

We are motivated to design and test a vocoder analyzer-synthesizer which adapts its time-frequency resolution rectangle to match the relative stationarity of different segments of the input speech. A digital implementation seems appropriate. The design of such an adaptive vocoder is discussed in the next chapter.

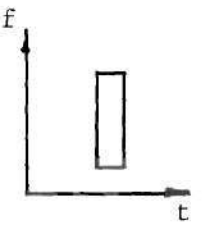
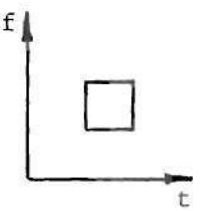
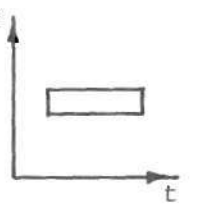
Type of Sound	Gross Spectrum Shape	Time Resolution	Frequency Resolution	Suggested Resolution Rectangle
"Short" (e.g., the Stop Consonant /t/)	Broadband and Smooth, Changing Rapidly	Important	Not Important	
"Intermediate" (e.g., a Voiced Consonant)	Changing	Relatively Important	Relatively Important	
"Long" (Sustained Vowels)	Changing Slowly	Not Important	Important	

Figure 20. Relative Importance of Time and Frequency Resolution

CHAPTER IV

THE ADAPTIVE HOMOMORPHIC VOCODER

Introduction

The experimental goal of this research effort is to design and implement a vocoder simulation system with which to test the utility of the adaptive analysis-synthesis strategy. The adaptive approach is an attempt to improve the short-time spectrum coding of the modern vocoder.

The motivation for the adaptive approach follows from the non-stationary nature of the speech signal. We seek to match the time and frequency resolution rectangle of the analyzer to the short-time "duration" and "bandwidth" of the signal.

The homomorphic vocoder is a natural choice for testing the adaptive approach because the time-frequency properties of the analyzer may be readily manipulated. Time resolution is controlled by the duration of the window function, and frequency resolution depends directly on the number of cepstrum coefficients transmitted.

In this chapter we motivate the choice of the homomorphic vocoder as a test platform for the adaptive approach. The design considerations are examined and the important parameters specified. The implementation of the experimental simulation system is described in detail. The chapter concludes with an outline of the vocoder software package.

The Homomorphic Vocoder -- A Natural Test Platform

We seek a vocoder system with time and frequency properties which may be easily manipulated. The simulation of such a system will serve as a test platform for examining the utility of the adaptive approach.

Consider the dual models of short-time spectrum representation illustrated in Figure 7 of Chapter II. The use of time sections, as generated from a bank of filters, is inappropriate to adaptive analysis, since modifying the number of filters on a short-time basis would introduce discrete "jumps" in time in a representation already discrete in frequency. Thus, one selects the frequency section approach. The short-time spectrum surface is represented by a sequence of sections $|S(t_i, f)|$ with variable spacing in time, each section retaining frequency resolution inversely related to the time interval it represents.

An obvious way to implement an adaptive frequency section strategy is with the short-time DFT discussed in Chapter II:

$$S_r(kF) = \sum_{n=0}^{N-1} w_i(nT)\ s(nT + A_r T)\ e^{-j2\pi nk/N} \qquad (4\text{-}1)$$

The subscript i is added to denote one of several adaptive modes. A_r is an integer which specifies the sample number at which the r^{th} analysis frame begins.

An example of the adaptive use of (4-1) is the following. Suppose $w_i(nT)$ is a Hanning window of duration $D_i = 10i$ ms, i = 1, 2, and 4. Let $A_r = 100r$, N = 512, and $T = 10^{-4}$ sec, so the window is propagated along the input signal in integer multiple steps of 10 ms. For a partic-

ular frame, say r_1, a frame decision is made, say $i = 2$, and $S_1(kF)$ is computed using a 20 ms window. The index r is incremented by i, so the next frame is $r = 3$. This example describes analysis with no overlap between adjacent frames.

We observe that in the example the sampling interval in each spectrum section is $F = \frac{1}{NT} \approx 20$ Hz. But this sampling interval does not indicate the frequency resolution inherent in each section, which depends on the inverse of the window function duration. For the Hanning window, the lobe bandwidth is $B_i = 4/D_i$ Hz. Thus, for the adaptive modes $i = 1$, 2, and 4, the short-time spectrum resolution cells are 10 ms $\times$ 400 Hz, 20 ms $\times$ 200 Hz, and 40 ms $\times$ 100 Hz respectively.

A spectrum analysis scheme employing the short-time DFT of (4-1) provides so-called "high resolution" spectrum sections. That is, no smoothing is introduced, and the only limitation on the frequency resolution obtained is caused by the window function. Such a spectrum resolves the individual harmonic components of a voiced sound, if the window spans several pitch periods.

The strategy of the vocoder is to transmit a pitch signal (which incorporates v/uv information) and a coding of the vocal tract spectrum. Thus, the high resolution spectrum generated by (4-1) is not an appropriate coding. Smoothing of the spectrum sections must be introduced to remove the pitch harmonic structure and reduce the dimensionality of the representation, to obtain an appropriate vocal tract coding.

One approach to coding the DFT sections is to average groups of frequency (amplitude) samples. But this approach would discretize the representation in frequency.

Consider the coding of the DFT sections employed in the homomorphic vocoder. The cepstrum (the inverse DFT of the logarithm of an amplitude section) is truncated to accomplish the smoothing of the spectrum, and the samples retained are the transmitted parameters. Thus, the cepstrum is a natural domain in which to manipulate the frequency resolution. The cepstrum in each frame is simply truncated to a number of samples proportional to the window function duration in that frame. This approach accomplishes the desired result, frequency resolution in the short-time spectrum coding inverse to the time resolution of the representation.

A simplified viewpoint of the cepstrum coding is to consider the cepstrum samples as a set of coefficients of the Fourier series expansion of the log spectrum. We retain many coefficients to describe the detailed shape of a sustained vowel spectrum, and few coefficients to represent the gross spectrum shape of a transitory consonant.

Thus, the homomorphic vocoder is naturally suited to the adaptive approach. We discuss the design of an adaptive homomorphic vocoder in the next section.

Design

The homomorphic vocoder was selected to implement the adaptive spectrum analysis strategy. A block diagram of an adaptive analyzer is shown in Figure 21. One conceives three parallel analyzers, with different window functions $w_i(nT)$. A decision is made for each analysis frame to determine the adaptive mode appropriate to that frame. After the

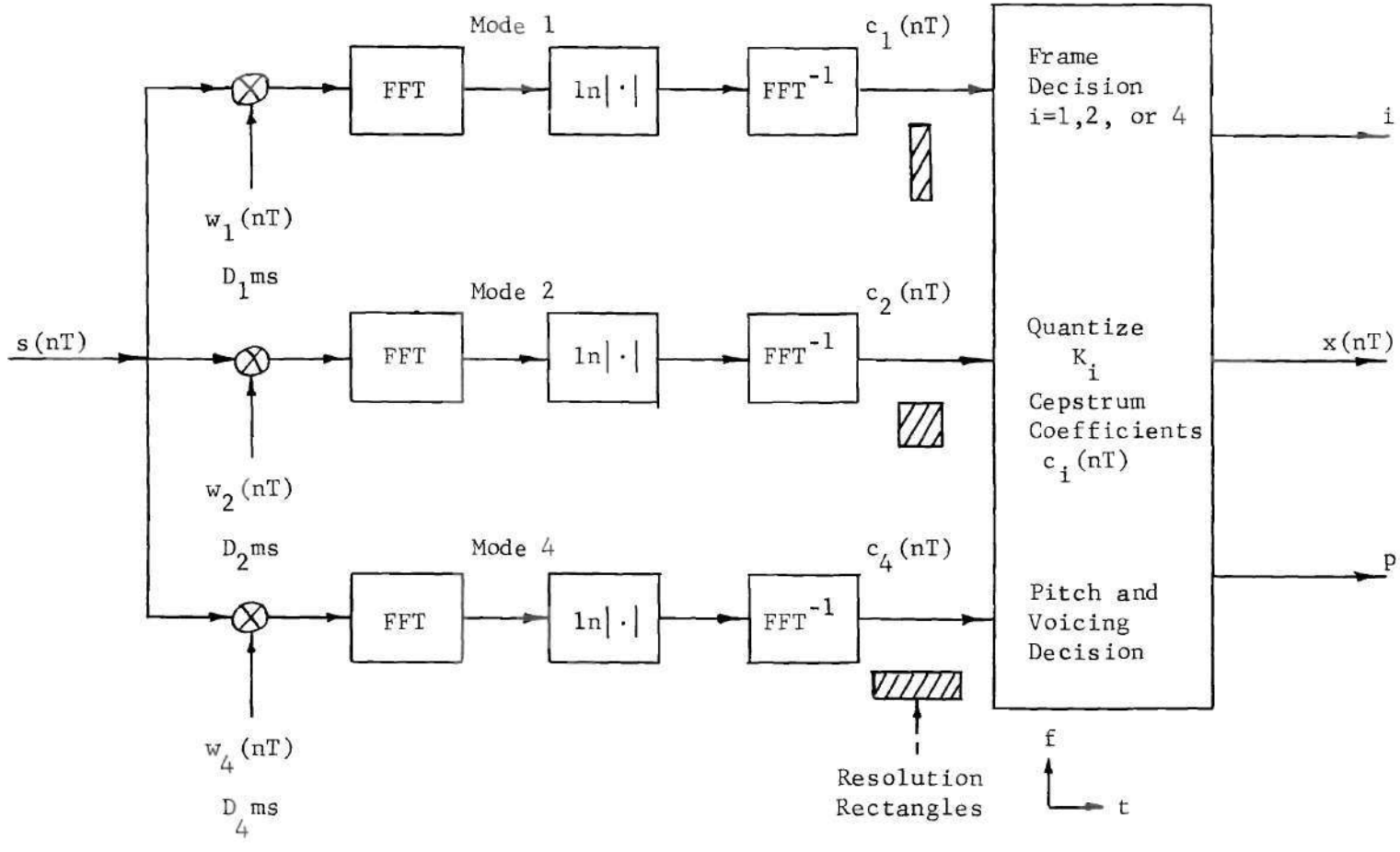


Figure 21. A Parallel Processing Model of the Adaptive Analyzer

"frame decision" is made the cepstrum parameters from the corresponding analyzer are selected for transmission. The synthesizer used with the adaptive analyzer may be identical to the conventional synthesizer, except that the convolution operation is extended over a time interval corresponding to the analysis frame.

A Frame Decision Strategy

In 1967 Gold and Rader described a channel vocoder experiment which employed a variable sampling rate for digital transmission of the channel signals, $x_n(t)$ [10]. The approach was to sample the spectrum signals rapidly when the spectrum was changing rapidly with time, and slowly otherwise. Notice that this strategy provides improved time resolution during transient features of the short-time spectrum, but retains uniform frequency resolution, thus presenting a changing bit rate signal to the digital channel.

Gold and Rader reported that the rate of a 2400 b/s vocoder may be reduced to about 1800 b/s with little appreciable degradation of the synthetic speech. This variable sampling rate experiment, and the philosophy of its conception, added motivation to the adaptive strategy.

Gold and Rader described the use of frame intervals of 15, 25, 30, 35, and 40 ms, the "mode" selected by comparing the "spectral derivative":

$$d(t) = \frac{\sum_{n=1}^{N} |\dot{x}_n(t)|}{\sum_{n=1}^{N} x_n(t)} \qquad (4\text{-}2)$$

to a set of thresholds. (The notation is that of Chapter II, $x_n(t)$ is the n^{th} spectrum channel signal.) The numerator of d(t) is the sum of magnitudes of the derivatives of the channel signals, while the denominator normalizes the measure to reduce its sensitivity to changes in speech volume.

The strategy of the spectral derivative seems an appropriate one for making the frame mode decisions in an adaptive processor. The channel signals used in (4-2) represent the smoothed spectrum section, since the channel filter bandwidths are too broad to resolve pitch harmonics. But a smoothed version of the spectrum section is not generated in the homomorphic analyzer.

A coding of the smoothed spectrum section that is available in the homomorphic analyzer is the low order cepstrum coefficients. A measure patterned after (4-2) sums the increments between successive cepstrums:

$$d_{ri} = \frac{\sum_{n=1}^{K_i} |c_{r+i}(nT) - c_r(nT)|}{K_i} \qquad (4\text{-}3)$$

in which we have decreased sensitivity to speech volume by excluding the c(o) coefficient in the summation. c(o) describes the "dc level" of a log-spectrum section. The summation is normalized by K_i -- the number of cepstrum coefficients transmitted in adaptive mode i.

The measure proposed in (4-3) may be interpreted as the "distance" between successive cepstrums in the linear space of cepstrum coefficients.

The decision of which analysis mode is appropriate for transmitting the "present" frame information is made by comparing d_{ri} to an experimentally determined threshold:

$$d_{ri} \underset{\text{try again}}{\overset{i}{\gtrless}} \alpha_i \qquad i = 1,2 \qquad (4\text{-}4)$$

The use of d_{ri} in the analyzer of Figure 21 is summarized as follows. Two successive cepstrums computed from the "short" window w_1 are used in (4-3), the resulting d_{r1} compared to threshold α_1, and if $d_{r1} > \alpha_1$ the spectrum surface is changing rapidly with time, so we choose mode 1 and transmit K_1 coefficients of the cepstrum. If $d_{r1} < \alpha_1$, we repeat the process with two cepstrums computed from w_2, and if $d_{r2} > \alpha_2$ choose mode 2. If $d_{r2} < \alpha_2$ mode 4 is selected. Observe that this decision strategy requires some fixed delay in the processor (additional to that inherent in the homomorphic analyzer) since one "future" cepstrum is needed in (4-3) to decide how to transmit the "present" frame.

The delay introduced with the cepstrum-distance decision-criterion would inhibit real-time operation of an adaptive vocoder, but is not important in the simulation to test the adaptive concept. A much more easily obtained measure could probably be devised.

Window Function Duration and the Frame Interval

The approach to implementing the adaptive analysis concept illustrated in Figure 21 is to alternate back and forth between two or three fixed modes as the nature of the signal changes. We consider in this

section what window durations and frame intervals are appropriate.

In his homomorphic vocoder, Oppenheim used a D = 40 ms Hanning window propagating in steps of MT = 20 ms [18]. Thus, the window in each frame overlaps half the previous window. The effect of the overlap is to smooth the short-time spectrum sections in time, increasing the correlation between successive sections. On the other hand, each section corresponds exactly to a sampled version of a time section, smoothed in frequency, of the short-time Fourier transform $|S(t_r, f)|$. Notice that the homomorphic analyzer has no smoothing in the time dimension except the "virtual smoothing" which obtains from the window function.

One argument in favor of the overlap is the following. Consider a time section of the short-time amplitude spectrum $|S(t, f_1)|$ which employs a Hanning window of duration D. The most rapid variations in time which can be produced in this section result from an input impulse train with period D. Such an input yields an output $|S(t, f_1)| = \cos \frac{2\pi t}{D}$. By the sampling theorem, since $|S(t, f_1)|$ may be considered to be band limited to 1/D Hz, the sampling rate should be 2/D, suggesting a frame interval of D/2. This sampling rate is sufficient to permit "ideal" (mathematical) interpolation of the amplitude spectrum surface in time. But using a one-dimensional processor in synthesis limits "ideal" interpolation to only one coordinate -- time or frequency. Such interpolation is achieved (under proper conditions) in the time dimension of a channel vocoder synthesizer and in the frequency dimension of a homomorphic synthesizer.

An assumption included in Oppenheim's development of the homomorphic deconvolution of speech is that the window function be approx-

imately constant over the duration of the impulse response v(t). Additional motivation for using a window function of 40 ms is that shorter windows inhibit the performance of cepstrum pitch detection [19].

Oppenheim employed linear interpolation between the synthesized impulse response functions of successive frames. Consider a transition region between stable vocal tract configurations. The transition would seem to be more nearly "tracked" in the linearly-interpolated synthesized-spectrum if the interpolation were accomplished between configurations represented by spectrums derived from time windows with little (or no) overlap.

It seems clear, however, that analysis with a 40 ms window combined with linear-interpolation in synthesis "smears" a "short" time event in the input speech to three 20 ms frames in the synthesized output. Furthermore, interpolation between impulse responses which represent vocal configurations on opposite sides of a discontinuity (such as from a voiced to unvoiced excitation) seems inappropriate.

We conclude that the window function durations and overlap must be selected experimentally. Window durations ranging from 10 to 51.2 ms appear appropriate for investigation.

The Analyzer

For simplicity in the vocoder simulation the frame decision and pitch detection functions are performed in a preliminary analysis run for each test sentence. Thus, the simulated vocoder analyzer has as inputs a set of frame and pitch decisions so that computer processor time is not required for these functions in successive vocoder runs of the same sentence.

The spectrum analysis functions of the homomorphic vocoder are retained in the simulation, and the analyzer parameters modified from frame to frame to achieve the desired adaptive action. The adaptive analyzer simulator is shown in Figure 22. The Hanning window function is employed because of its use in previously reported homomorphic vocoder designs [18,21]. (One might consider the use of a prolate spheroidal wave function window [22].)

The features of the analyzer necessary to completely describe its operation are outlined as follows:

1. The rule used to compute the starting address of the r^{th} frame, e.g., $A_r = 100r$, increment r in steps of i.

2. The window function duration for each adaptive mode, e.g., $D_i = 12.8i$ ms.

3. The rule for computing the FFT radix to be used, e.g., $N_i = 128i$.

4. The number of cepstrum coefficients transmitted in mode i, e.g., $K_i = 10i$.

5. The quantizer characteristic, e.g., quantize the known range of cepstrum coefficients uniformly to Q bits (2^Q levels).

The Synthesizer

The adaptive synthesizer is essentially equivalent to that reported by Oppenheim, discussed in Chapter II. The only required change is to limit the convolution operation to the frame interval. The synthesizer is shown in Figure 23.

The excitation generator produces a train of unit pulses spaced at the pitch period for voiced frames, and a train of pulses of random polarity spaced 1 ms for unvoiced frames.

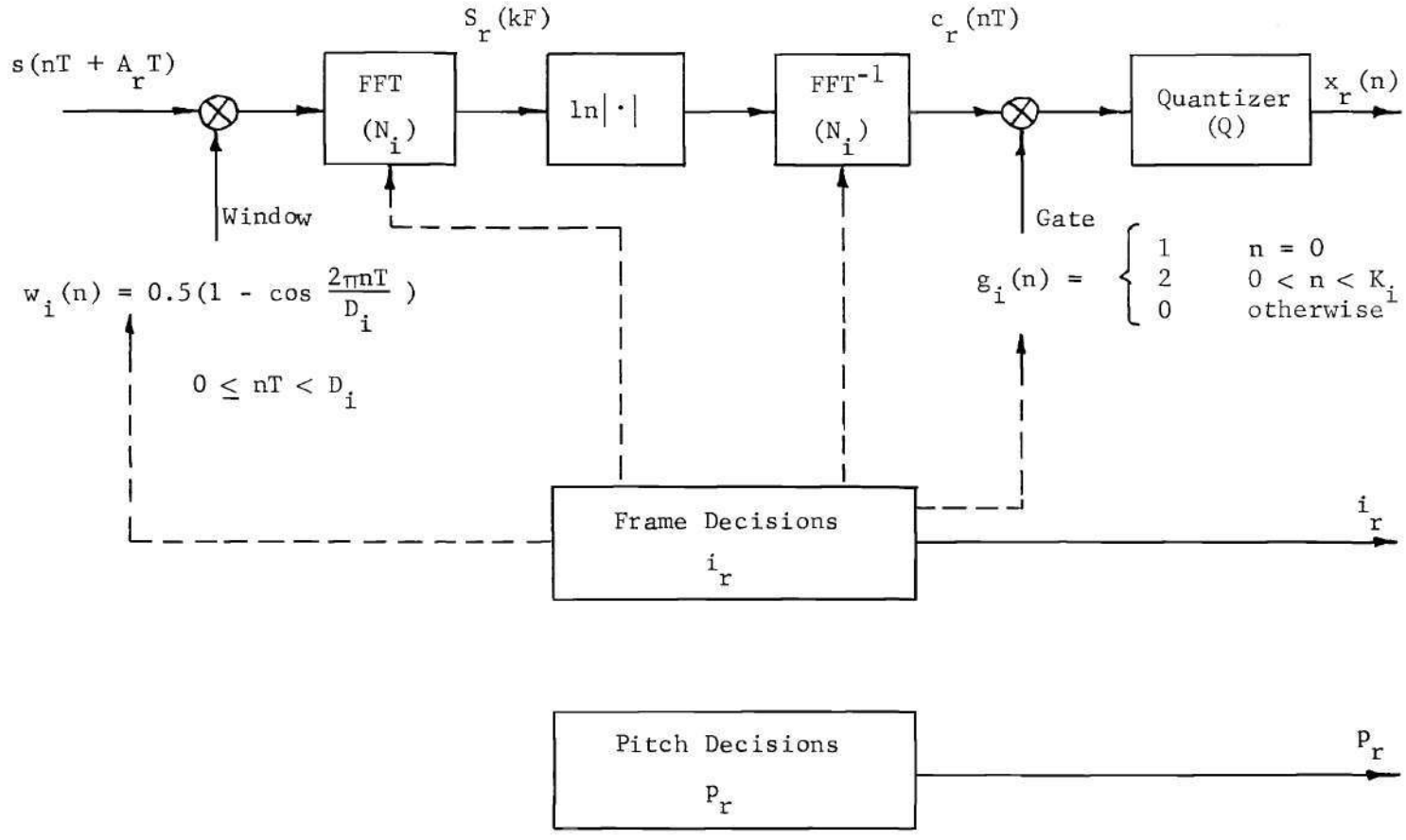


Figure 22. The Adaptive Analyzer

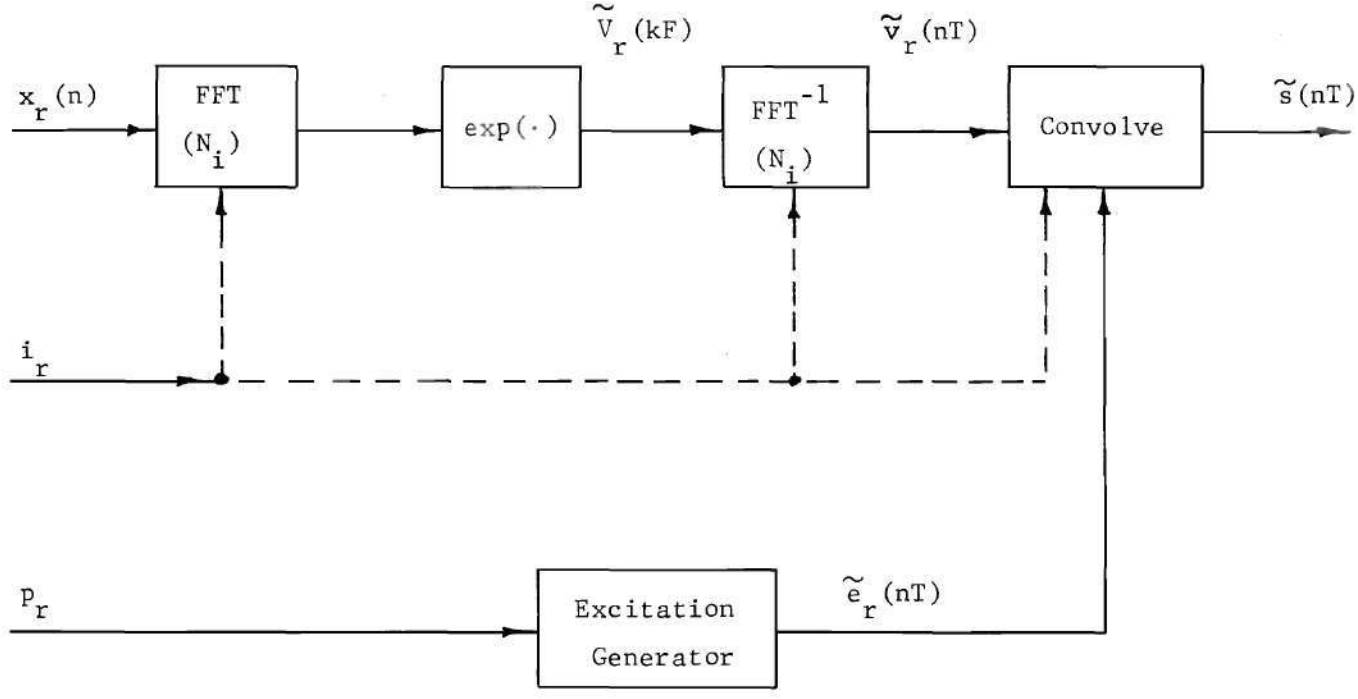


Figure 23. The Adaptive Synthesizer

The convolution operation retains the linear interpolation feature between impulse response functions of adjacent frames. For example the impulse response used t_1 ms into the r^{th} frame is

$$v(nT) = \left(1 - \frac{t_1}{20}\right) v_{r-2}(nT) + \left(\frac{t_1}{20}\right) v_r(nT) \qquad (4\text{-}5)$$

where we have assumed interpolation between 20 ms frames. The interpolation is performed only between voiced frames, and in the event of adjacent frames of different duration, the interpolation is completed in an interval equal to the duration of the shorter frame.

The Experimental Vocoder Simulation System

The stages of the simulation system are listed in Figure 24.

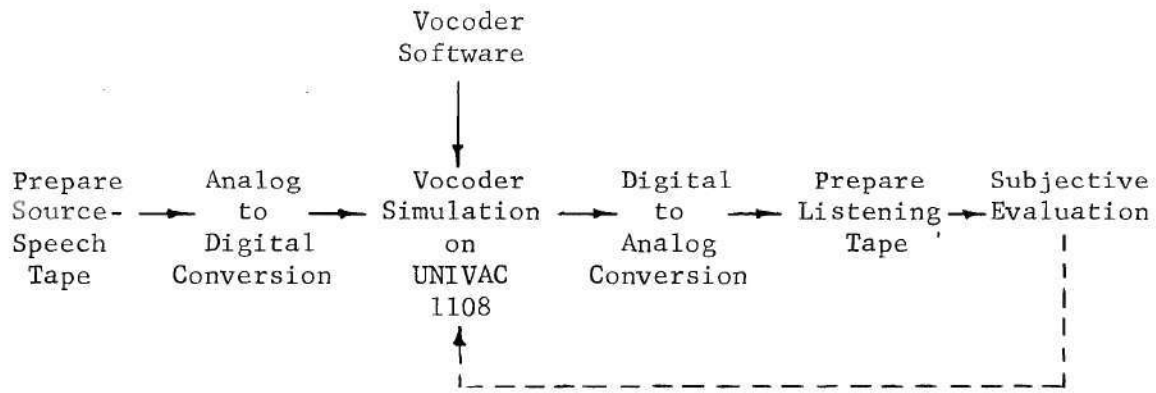


Figure 24. Stages of the Simulation System

A source-speech tape was produced with fourteen sentences spoken twice by each of six talkers, three male and three female. A Shure Model 51 Dynamic Microphone was positioned in a quiet room designed for speech recording and listening tests. The quiet room is a facility of the

Communications Branch, Electronics Division, of the Georgia Institute of Technology Engineering Experiment Station. Recording from the quiet room was accomplished on an Ampex Model 351 Tape Recorder using 1/4-inch tape at 7-1/2 in/s. The sentences recorded by each talker are the following:

1. The waves looked threatening.
2. He took a walk every morning.
3. He gave me a corsage.
4. Your shouting was inexcusable.
5. The sixth grade had a picnic.
6. Your gift is a birthday cake.
7. We met at the junction.
8. They march in precision.
9. You've been measuring the width.
10. Your jumping thrilled him.
11. Your jingle was first.
12. Give the cashbox to me.
13. It's easy to tell the depth of a well.
14. We were away a year ago.

The analog-to-digital conversion processing is shown in Figure 25. The source tape was played on an Ampex Model SP-300 Tape Recorder through an EAI Model TR20 Analog Computer which performs the following operations:

1. Signal amplification to 2.0 volts peak-to-peak
2. Low-pass filter at 4 kHz, 6 db/octave
3. Clip at 3 volts peak-to-peak

The resulting speech signal was then filtered by two, 2-section Krohn-

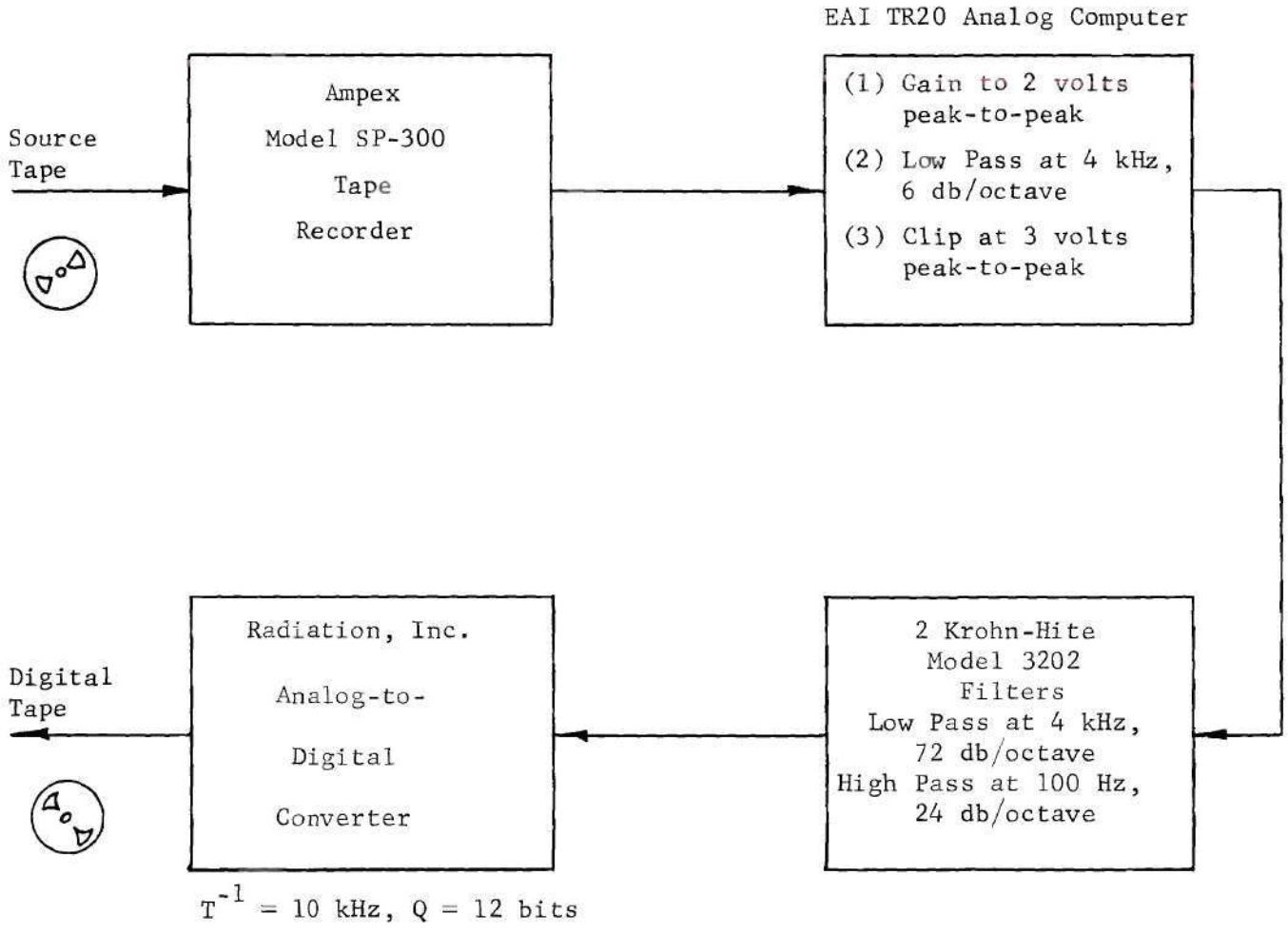


Figure 25. Analog-to-Digital Conversion

Hite Model 3202 Filters, one high-pass section providing 24 db/octave attenuation below 100 Hz, and three low-pass sections providing 72 db/octave attenuation above 4 kHz. Each filter section has a maximally-flat three-pole-Butterworth system function. The filter output drives a Radiation Inc Analog-to-Digital Converter which samples the analog input at 10 kHz, quantizes the samples uniformly to 12 bits (4096 levels), and writes the 12 bit data words on IBM-compatible digital tape at 556 bits per inch.

The principal facility used in the simulation is the UNIVAC 1108. The vocoder software package is organized in 39 subroutines written in FORTRAN V. The UNIVAC operations are indicated in Figure 26. A sentence from the digital tape is read into core memory, the vocoder operations performed, and outputs obtained by line printer, a CALCOMP Digital Incremental Plotter, and paper tape. The CALCOMP Plotter displays the waveforms computed in the vocoder, as well as the synthesized speech. Samples of the synthesized speech are quantized to 10 bits and punched on paper tape. The simulation operates at a processor time of about 80 times real-time. Any of the data or waveforms in each frame can be called out and plotted on the CALCOMP plotter.

Digital-to-analog conversion is accomplished on a DEC PDP-8 computer, as illustrated in Figure 27. The paper tape of the digital synthesized speech is read by a high speed reader into the RF08 Disk Memory. The disk is then "played" through core memory (under a data-break transfer) and a digital-to-analog converter (interrupt driven by a programmable clock). The D/A output is band-pass filtered to 100 Hz -- 4 kHz by

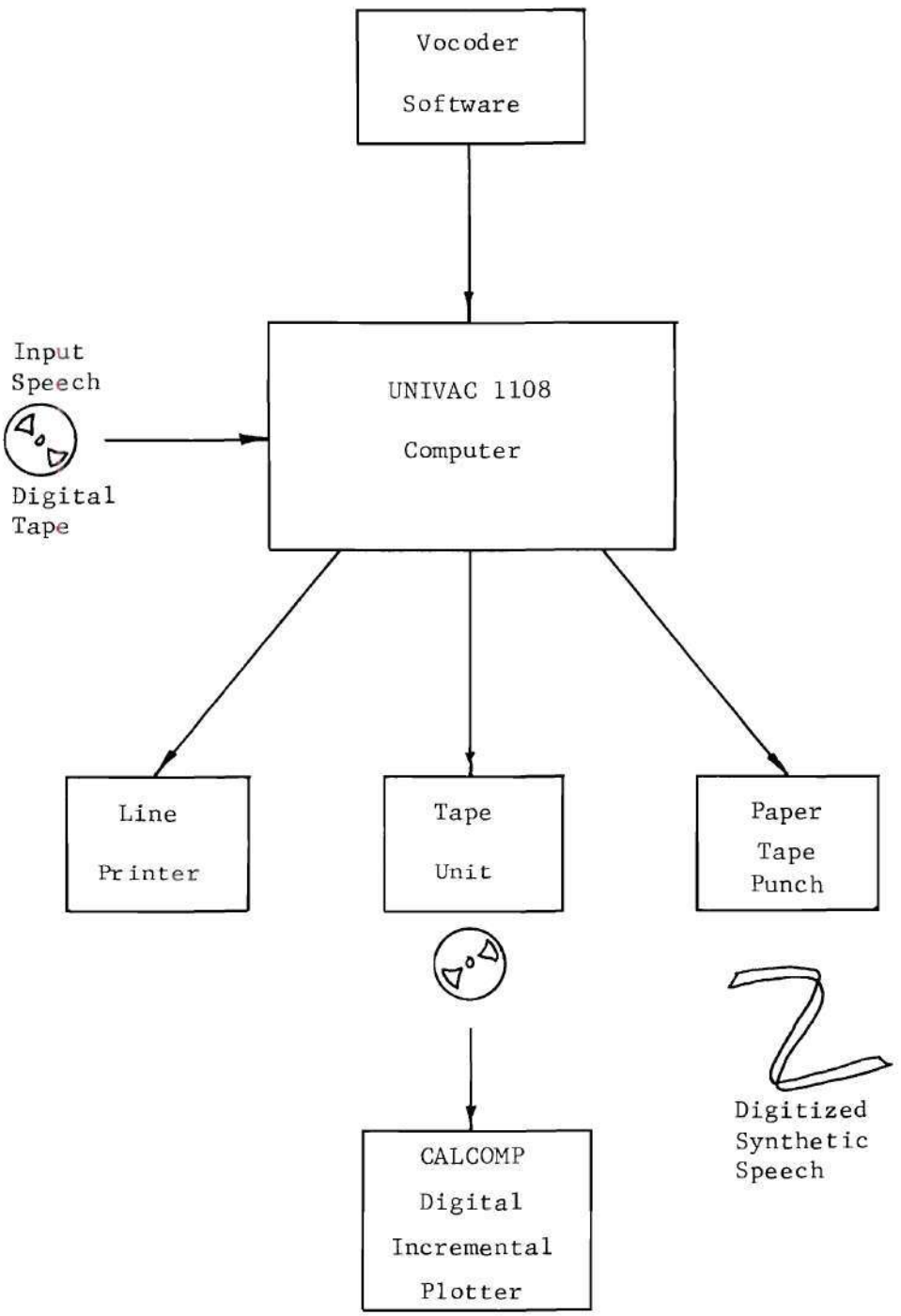


Figure 26. The Vocoder Simulation

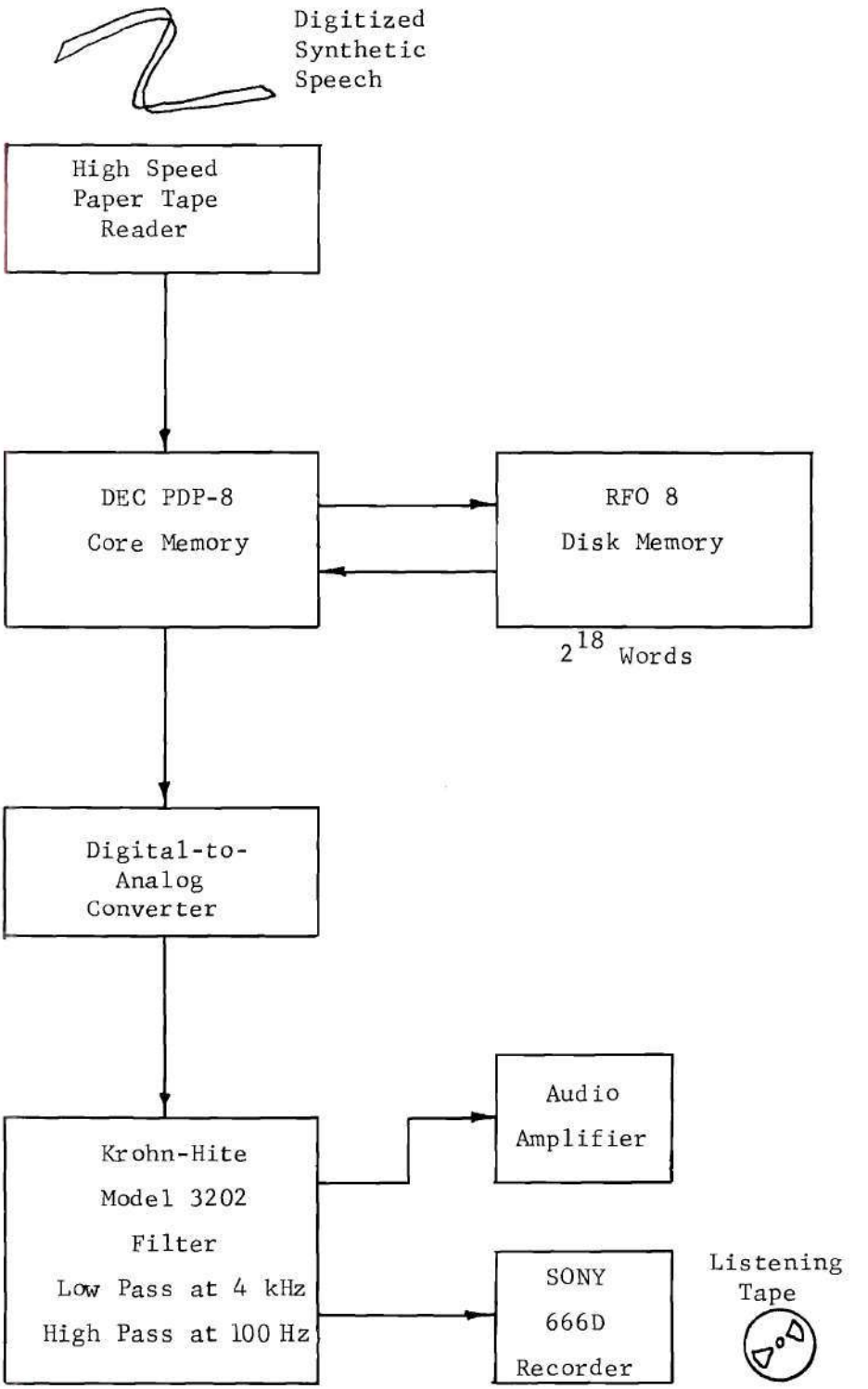


Figure 27. Digital-to-Analog Conversion

a Krohn-Hite Model 3202 filter. The resulting analog speech is recorded on a SONY 666D tape recorder.

The resulting listening tape of vocoded speech is played on the AMPEX 351 for subjective listening tests and the preparation of Sonagrams on a KAY Model 7029A Sonagraph.

The Vocoder Software

The vocoder software library was organized in 39 subroutines written in FORTRAN V. The routines were evolved over 280 runs on the UNIVAC 1108. Vocoded speech output was obtained on approximately 85 of these computer runs.

The FFT Algorithm

The heart of the software package is the FFT algorithm. The algorithm selected for use in the vocoder simulation is one due to Uhrich which leads to a particularly simple and compact FORTRAN program [37]. Most FFT algorithms offer the advantage of computation in-place (that is, the output is returned in the input array, without the need for buffer storage), but require a sorting of the output array, which is rendered in bit-reversed index order. The simplicity of Uhrich's algorithm obtains from its use of twice as much core memory, and its naturally ordered output which requires no sorting. Signal flow graphs of various FFT algorithms, including that of Uhrich's algorithm, are given in Gold and Rader [8].

One modification to Unrich's algorithm was incorporated. At every call to the FFT subroutine, the array of complex coefficients $\{e^{\pm 2\pi k/N}\}$

$k = 0, 1, \ldots, N-1$ is required. Since a vocoder run requires several hundred calls to the FFT, the array of coefficients is computed by subroutine COMPW, the coefficients stored in the common array W available on demand by the FFT. The program listing of COMPW is shown in Figure 28. The FFT algorithm, named HAMFT, is shown in Figure 29.

Nonlinear Operations

The following subroutines perform the nonlinear operations required in the vocoder simulation:

1. CMLOG computes the natural logarithm of a real input array.
2. CXEXP computes the complex exponential of a complex input array.
3. MAGY computes the magnitude of a complex input array.
4. TRUNC truncates a real input array.
5. QTIZE quantizes a real input array uniformly to a specified number of levels over a fixed range.
6. QTZ1 quantizes a real input array uniformly to a specified number of levels over an adjustable range.
7. WINDO multiplies a real input array by a Hanning window sequence, for windows containing a number of samples that equals an integer power of two.
8. WINDU multiplies a real input array by a Hanning window sequence of arbitrary length.

Array Loading

The following subroutines perform various array loading operations:

1. LOADA loads one integer array into another.
2. LOADB loads an integer array into a real array.

```
1*            SUBROUTINE COMPW
2*      C     THIS SUBROUTINE COMPUTES COSINES AND SINES ON THE INTERVAL (0,PI)
3*      C               W(1,K) = COS(2*PI*(K-1)/4096)
4*      C               W(2,K) = SIN(2*PI*(K-1)/4096)
5*            DOUBLE PRECISION PI2ON,B
6*            COMMON W(2,2048)
7*            PI2ON=3.1415926535897932D0*2.0D0/4096.0D0
8*            DO 116 K=2,1024
9*            B=PI2ON*(K-1)
10*           S=DCOS(B)
11*           W(1,K)=S
12*           W(1,2050-K)=-S
13*           W(2,1026-K)=S
14*           W(2,K+1024)=S
15*       116 CONTINUE
16*           W(1,1)=1.0
17*           W(1,1025)=0.0
18*           W(2,1)=0.0
19*           W(2,1025)=1.0
20*           END
```

Figure 28. Subroutine COMPW

```
1*            SUBROUTINE HAMFT(Y,Z,M,N,SIGN)
2*      C     THIS SUBROUTINE COMPUTES THE DISCRETE FOURIER TRANSFORM OF THE COMPLEX
3*      C       INPUT SEQUENCE Y(M),  WHERE M=2**N.  IF SIGN=-1 THE DIRECT TRANSFORM
4*      C       IS COMPUTED.  IF SIGN=+1 THE INVERSE TRANSFORM IS COMPUTED.  THIS
5*      C       TRANSFORM IS ADAPTED FROM THAT OF M. L. UHRICH, IEEE TRANS AU-17,
6*      C       JUNE 1969.
7*            INTEGER M,N,SIGN
8*            M2=M/2
9*            COMPLEX Y(M)
10*           COMPLEX Z(M)
11*           COMMON W(2,2048)
12*           COMPLEX WK
13*           KSTEP=2**(12-N)
14*           DO 3 J=1,N
15*           NO2J=2**(N-J)
16*           NI=2**(J-1)
17*           DO 2 I=1,NI
18*           K=(I-1)*NO2J
19*           WKI=SIGN*W(2,K*KSTEP+1)
20*           WK=CMPLX(W(1,K*KSTEP+1),WKI)
21*           DO 2 L=1,NO2J
22*           I0=L+K
23*           I1=I0+K
24*           I2=I1+NO2J
25*           I3=I0+M2
26*           Z(I0)=Y(I1)+WK*Y(I2)
27*           Z(I3)=Y(I1)-WK*Y(I2)
28*         2 CONTINUE
29*           DO 3 L=1,M
30*         3 Y(L)=Z(L)
31*           IF (SIGN.LT.0) RETURN
32*           DO 4 L=1,M
33*         4 Y(L)=Y(L)/M
34*           RETURN
35*           END
```

Figure 29. Subroutine HAMFT

3. LOADC loads a real array into a complex array, with adjustable starting points.

4. LOADR loads a specified portion of one real array into a specified portion of another.

5. LOADX loads one real array into another, with the option to repeat the input periodically if the output array is larger.

6. CLOAD loads one complex array into another.

7. YCINX loads the real part of a complex array into a real array.

8. YINCX loads a real array into a complex array.

9. ZERO sets the elements of a real array to zero.

Digital Tape Operations

This group of subroutines reads the digital source tapes produced on the Radiation, Inc. Analog-to-Digital Converter:

1. CKTAP reads and prints out the identification and calibrate records at the beginning of every A/D tape.

2. RDTAP reads a specified number of records into an integer array, starting at the desired address on tape.

3. RECRD sorts the record numbers and data words obtained by RDTAP into a new integer array.

4. PRREC prints the integer array output from RECRD.

5. RDAPR combines the routines listed above to fill a real array with a specified number of A/D tape records, with optional printed output.

CALCOMP Plotter Routines

These subroutines provide flexibility in plotting the waveforms encountered in the vocoder:

1. PLUT1 plots one "large" real array, e.g., a synthesized sentence.

2. PLUT7 plots a real array of specified dimension on axes of adjustable size. Successive calls to this routine produce "columns" of four small plots.

3. PLUT8 plots one "large" array compactly using routine PLUT7.

4. RANGE examines a real array to determine its minimum and maximum in order to normalize an array to be plotted.

Data Input/Output Routines

These subroutines read data input from punched cards and provide data output by line printer and paper tape:

1. FILLP reads punched cards to input the random binary sequence used in unvoiced synthesis.

2. FILLIP reads punched cards to input the pitch decisions to be used in synthesis of a speech utterance.

3. FILLIW reads punched cards to input the adaptive mode frame decisions used in adaptive synthesis.

4. PRARR prints a real array of specified dimension compactly by line printer.

5. PRTP1 provides paper tape output of a large array of speech samples. The array is quantized to 10 bits, and each data word punched on paper tape in two 6 bit characters, coded in 2^S-complement form.

Vocoder Operations

These routines perform the operations required in the homomorphic vocoder:

1. PITCH examines the high-quefrency region of the cepstrum for a pitch peak.

2. VUV makes a voiced/unvoiced decision by comparing the relative energy in high and low frequency regions of the amplitude spectrum, and by the presence or absence of a prominent cepstrum peak.

3. ANALY performs the vocoder analysis operations for one frame.

4. CHANL performs the quantization of the cepstrum parameters, and simulates the coding of a communications channel.

5. SYNTH performs the minimum-phase synthesis operations to produce a synthesized impulse response for each vocoder frame.

6. CONVO performs the excitation generation and convolution operations of the synthesizer, including linear interpolation between successive impulse responses.

Summary

In this chapter the design and implementation of an adaptive homomorphic-vocoder simulation system was described. The homomorphic vocoder was selected to test the adaptive approach because of the simple, direct control that may be exercised over the time and frequency resolution characteristics of the processor. Only minor modifications of the homomorphic processor were required to achieve the adaptive action.

The vocoder simulation system is composed of an A/D Converter, a large-scale digital computer, a package of vocoder software , and a small computer which performs the D/A Converter function. The results obtained with this experimental system are discussed in the next chapter.

CHAPTER V

RESULTS

Introduction

The experimental results obtained with the vocoder simulation system are described in this chapter. The pitch-detection and adaptive-frame-decision functions were performed in preliminary analysis runs. The conventional homomorphic vocoder was operated with various window-function durations and frame intervals. The adaptive vocoder was simulated in several configurations and the resulting synthesized speech evaluated in subjective listening tests.

The results reported in this chapter suggest that the adaptive strategy has potential for improving the "quality" of vocoded speech.

Preliminary Analysis

Four sentences were selected as test inputs to the vocoder simulation. The sentences are the following:

1. Your gift is a birthday cake.
2. The sixth grade had a picnic.
3. We were away a year ago.
4. It's easy to tell the depth of a well.

The first sentence was spoken by a female talker, the other three by males. Sentences 1 and 4 were selected because of the many rapid transitions and stop consonants they contain. Such sentences present the

greatest challenge to the time resolution properties of a vocoder.

The second sentence is also characterized by rapid transitions. Sentence 3 is a test utterance used in formant analysis-synthesis work at the Bell Telephone Laboratories [2, 12]. The sentence is composed of voiced, non-nasal phonemes and is characterized by "slow" transitions. All four sentences were spoken at a rapid, conversational rate, with durations of 1.4, 1.2, 1.1, and 1.6 seconds respectively.

The Sonagrams of the first three test sentences are pictured in Figure 9 of Chapter II. These Sonagrams were prepared from recorded versions of the original sentences that had passed through the simulation system. That is, each sentence was A/D converted, quantized to 12 bits, read into the UNIVAC 1108, quantized to 10 bits on paper tape, and "played" through the D/A converter on the PDP-8.

Pitch Detection

The pitch detection function was performed in a preliminary analysis run of the homomorphic analyzer. Analysis was accomplished with a window function of relatively long duration, either 40 or 51.2 ms, with a frame interval of 10 ms. The PITCH algorithm was employed to pick the peak in the high quefrency region (4-12.8 ms) of each cepstrum.

CALCOMP plots of the cepstrums for successive frames were obtained to aid in the pitch detection process. The cepstrum plots for three frames of the "we" in Sentence 3 are shown in Figure 30.

To identify unvoiced frames, subroutine VUV was used to compute the relative energy in the high frequency (2500-3500 Hz) and low frequency (100-1100 Hz) regions to provide an indication of voicing. This

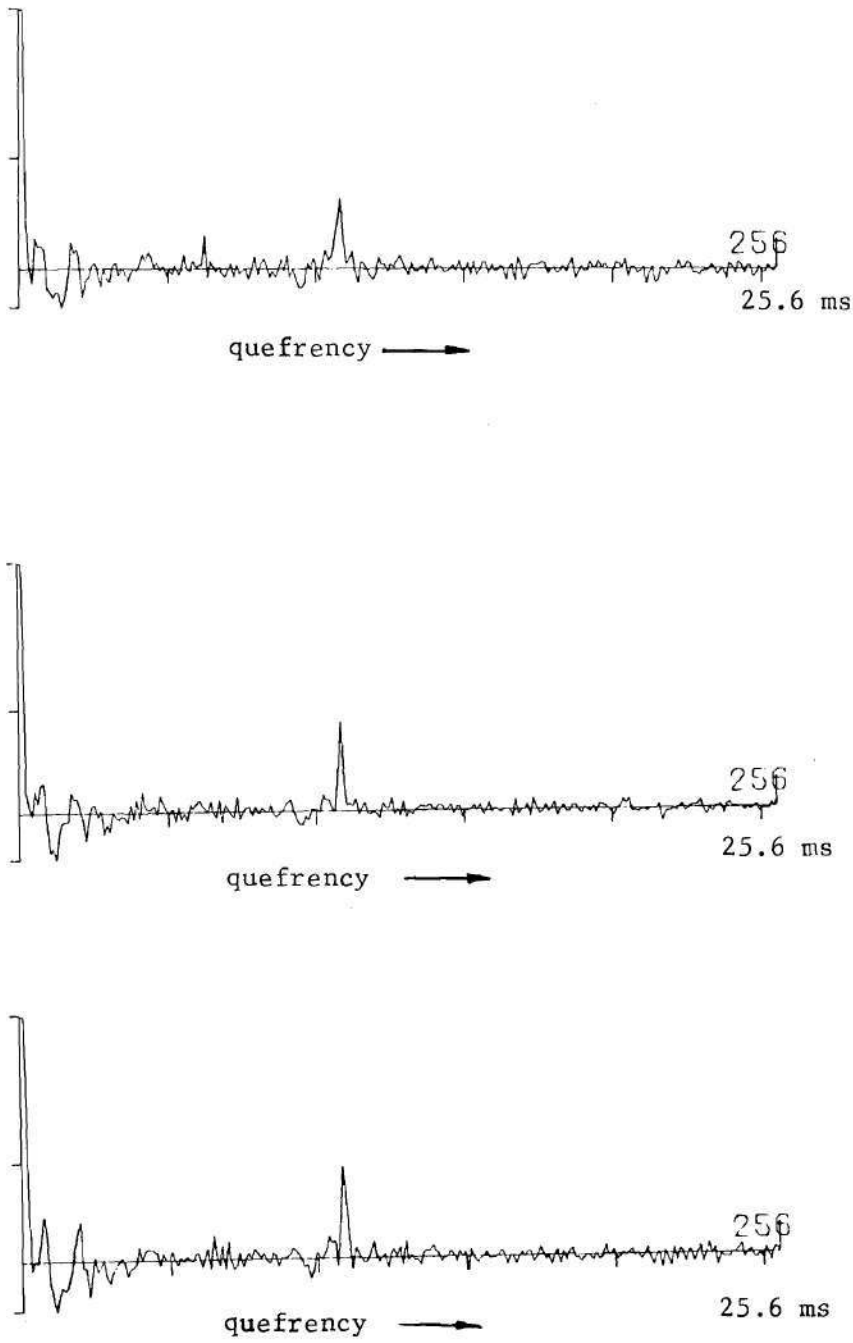


Figure 30. The Pitch Peak in the Cepstrum

energy measure and the presence or absence of a cepstrum peak were used in making voicing decisions.

Pitch detection by simple peak picking in the cepstrum was generally satisfactory except at the junction of voiced/unvoiced transitions. Pitch errors were corrected by hand editing. One limitation of cepstrum pitch detection is that the analysis window must be long enough to include about three (or more) pitch periods. Otherwise the computed amplitude spectrum may have insufficient frequency resolution to produce a prominent pitch peak in the cepstrum. A 25.6 ms window was long enough for pitch detection of Sentence 1 (spoken by a female) in which the pitch periods encountered ranged from 4.8 to 6.5 ms. But the 25.6 ms window was too short for Sentence 3, with pitch periods from 6.9 to 12.5 ms.

Thus, the pitch detection function was accomplished by a preliminary analysis, with corrections made after inspection of waveform plots and Sonagrams. The resulting pitch contour for Sentence 1 is plotted in Figure 31. Unvoiced intervals are indicated by a pitch of 1 ms.

Adaptive Frame Decisions

The frame decisions which control the adaptive operation of the vocoder analyzer were made in a preliminary analysis run for each test sentence. The cepstrum distance measure d_{ri} described in equation (4-3) was the basis for the frame decisions.

A cepstrum distance contour for Sentence 1 is shown in Figure 32. This contour was derived with a 25.6 ms analysis window propagating in steps of 20 ms. In this example, frames whose distance measure was less than 0.2 were considered as candidates for analysis with a "long" window.

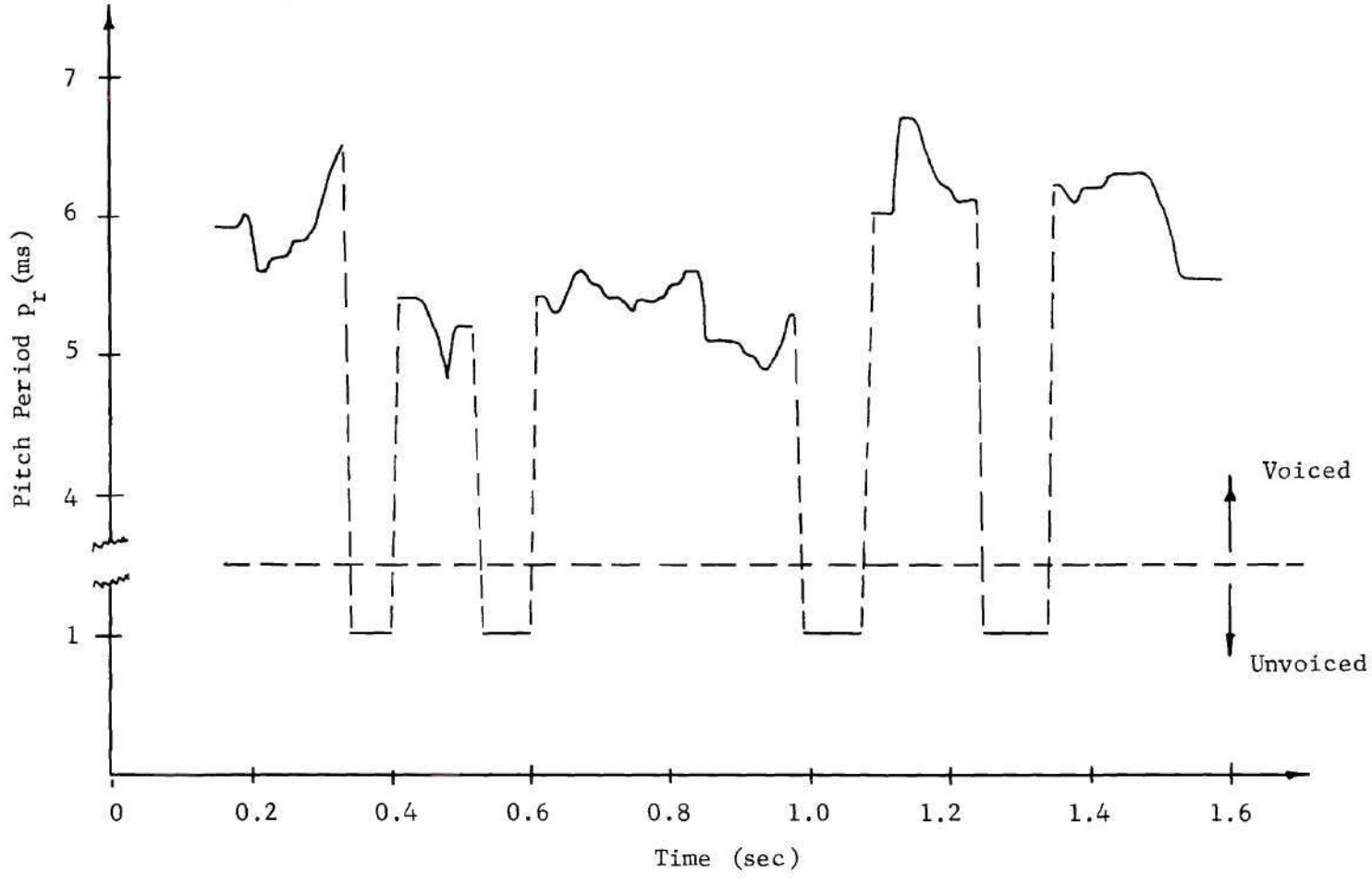


Figure 31. The Pitch Contour for Sentence 1

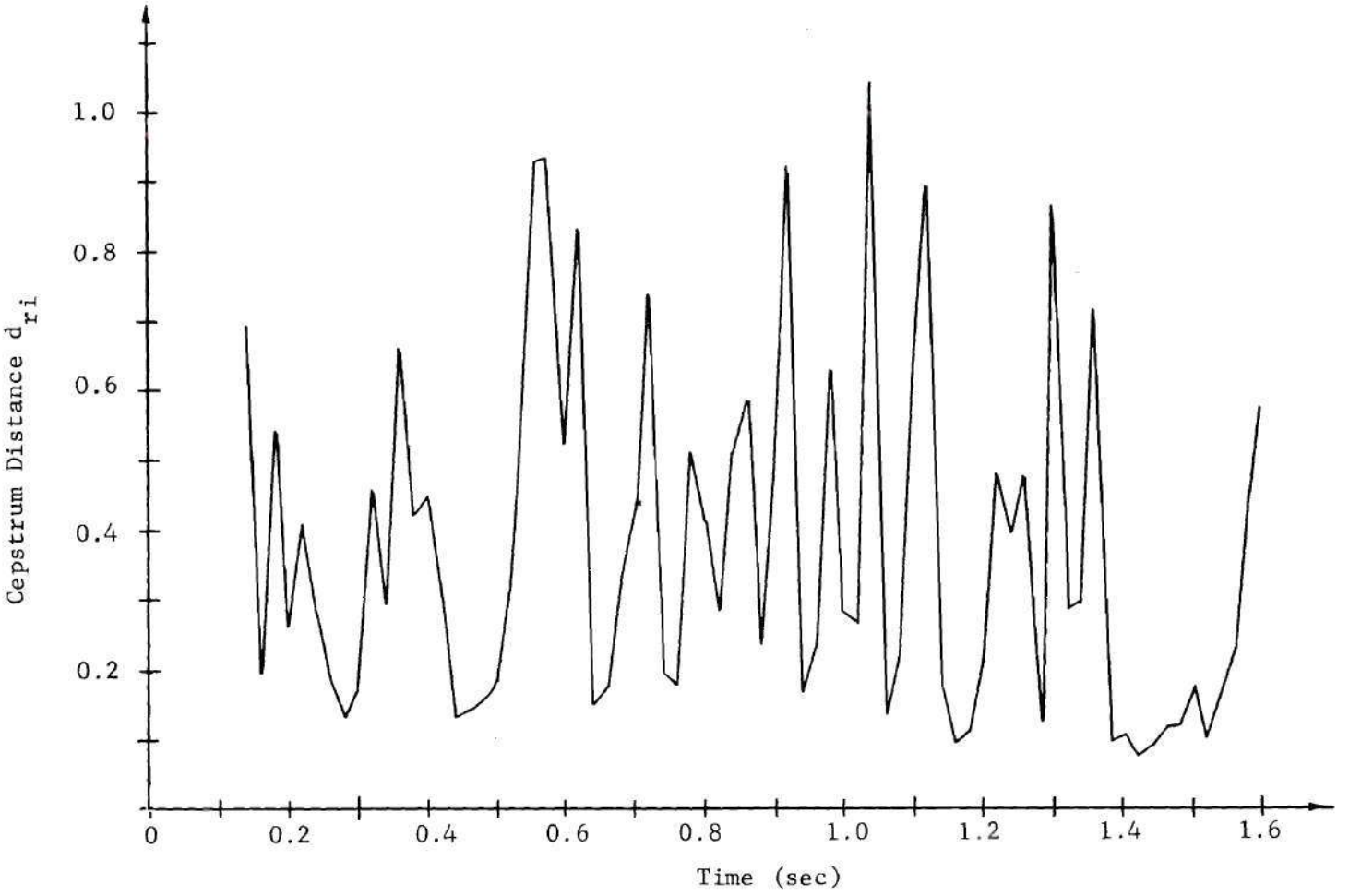


Figure 32. A Cepstrum Distance Contour

The cepstrum distance contours and an inspection of the speech waveform were used to make the frame decisions.

The pitch and frame decisions were coded as integer arrays and provided as inputs to the vocoder simulator on the UNIVAC 1108.

The Conventional Homomorphic Vocoder

Several experiments were conducted with the vocoder operating in the (non-adaptive) conventional mode. The purpose of these experiments was to compare the "quality" of speech synthesized from analysis with window functions of different durations.

Experiment 1

In this early experiment the conventional vocoder processed test Sentence 4. The sentence was processed three times, with analysis windows of duration 12.8, 25.6, and 51.2 ms respectively. The window was propagated in uniform steps of 25.6 ms, the cepstrum truncated to 40 coefficients without quantization, and minimum-phase synthesis accomplished with linear interpolation incorporated between voiced frames.

In an informal subjective listening test the sentence processed with the 51.2 ms window was judged to be the poorest of the three. Opinions were mixed in judging the 12.8 and 25.6 ms window runs, although a slight preference was indicated for the 25.6 ms window.

Experiment 2

In this experiment test Sentence 1 was the input to the conventional processor. The two parts of the experiment differed only in the window duration (D) and frame interval (FI) used:

Part	D(ms)	FI(ms)
1	25.6	20
2	12.8	10

The cepstrum was truncated without quantization to 24 coefficients.

In the listening evaluation of this experiment, a preference was indicated for Part 2. Notice that in Part 2, twice as many coefficients per unit time were "transmitted" in the simulated channel. Part 2 incorporates more time resolution in the short-time spectrum coding than does Part 1.

Experiment 3

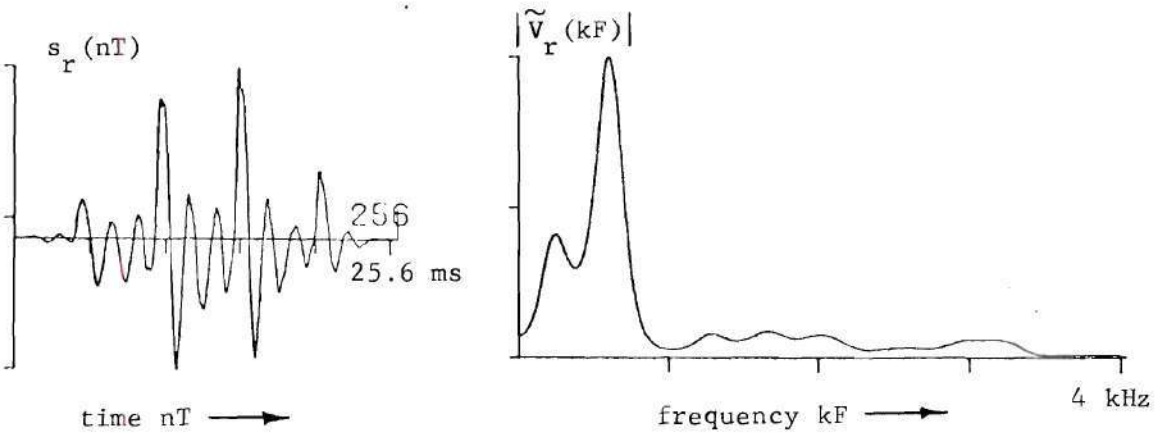
In this experiment Sentence 1 was processed three times by the conventional vocoder with the same window durations used in Experiment 1:

Part	D(ms)
1	25.6
2	12.8
3	51.2

The frame interval was constant at 20 ms, and the cepstrum was truncated without quantization to 30 coefficients.

As in Experiment 1, the 51.2 ms window yielded the poorest results. No preference was indicated between the shorter windows.

The vocoder waveform plots for one analysis-synthesis frame are shown in Figures 33 and 34, for Parts 1 and 2 respectively. Plotted in each figure are the windowed speech $s_r(nT)$, the cepstrum $c_r(nT)$, the



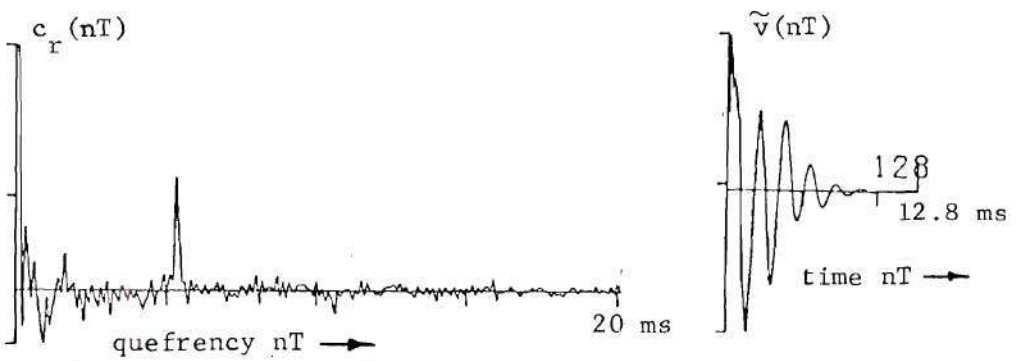


Figure 33. Vocoder Waveforms - Part 1 of Experiment 3

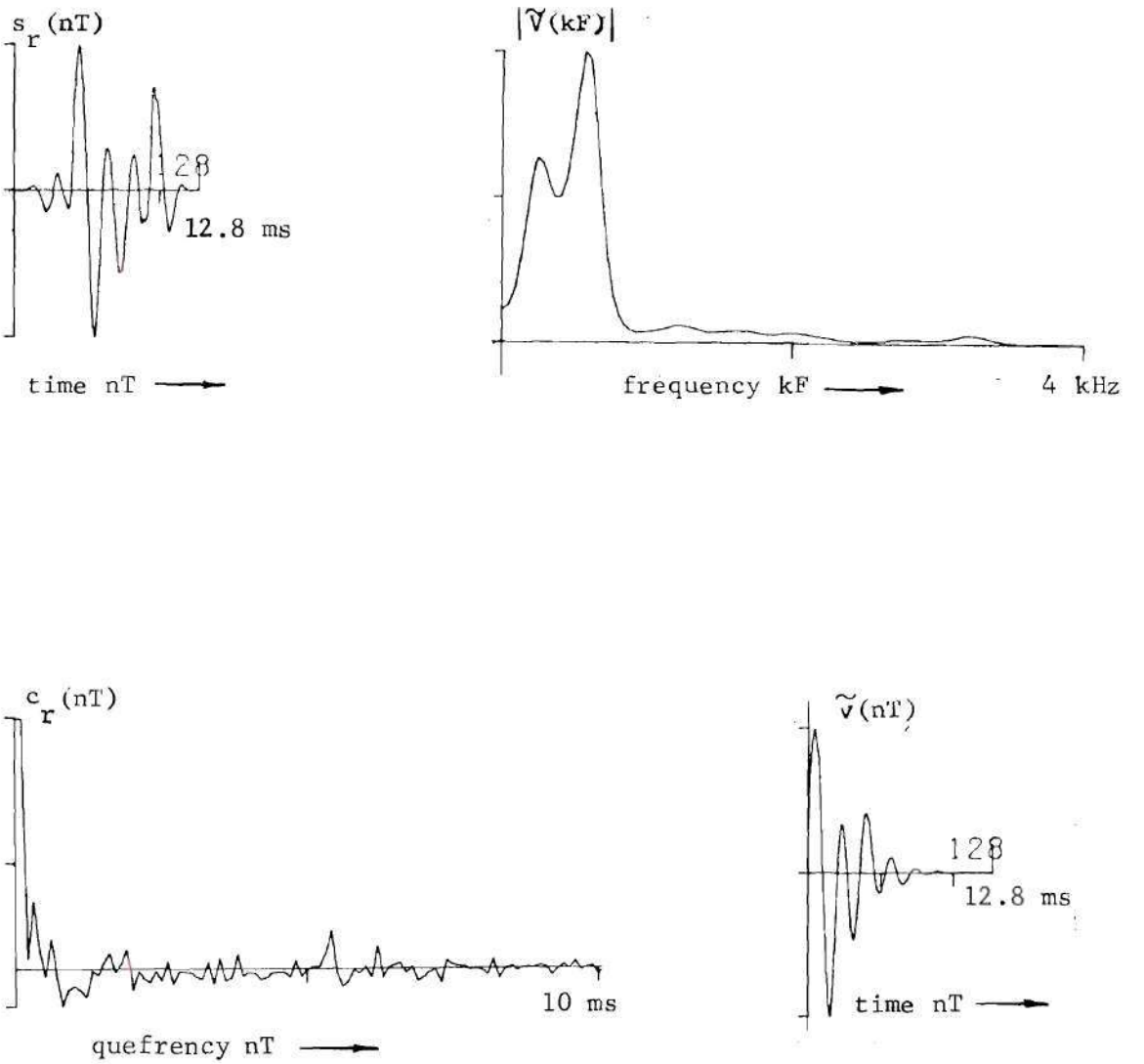


Figure 34. Vocoder Waveforms - Part 2 of Experiment 3

vocal tract spectrum $|\tilde{V}_r(kF)|$, and the synthesized impulse response $\tilde{v}(nT)$. The frame illustrated in the figures is r = 73, which begins 720 ms into the test utterance. This frame represents a portion of the phoneme /ʌ/ in "gift is a." The measured pitch period for this frame was 5.3 ms, so about 2.4 periods fall into the analysis window in Part 2.

Notice the close agreement of the impulse response functions obtained in Parts 1 and 2. The 12.8 ms window of Part 2 violates the assumption that the window is approximately constant over one pitch period, but a satisfactory deconvolution is clearly achieved, and satisfactory synthesized speech results.

This result suggests an interesting possibility. Notice that when the 12.8 ms window is used with a frame interval of 20 ms almost half of the original speech waveform is never "seen" by the analyzer. But satisfactory synthesized speech results. The complexity of the homomorphic vocoder is dominated by the FFT operations. The FFT complexity is proportional to $N \log_2 N$, where N is the length of the input sequence the FFT will accommodate. Thus, a complexity reduction by a factor of about 5 seems possible using the short window (12.8 ms, $128 = 2^7$ point FFT) rather than a long one (e.g., 40 ms, $512 = 2^9$ point FFT).

Conclusion

The tentative conclusion from the results of these three experiments with the conventional homomorphic vocoder is that 12.8 and 25.6 ms window functions yield synthesized speech of higher quality than does a 51.2 ms window.

The Adaptive Vocoder

The results of the adaptive homomorphic vocoder experiments are described in this section. The adaptive vocoder was implemented in the manner described in Chapter IV. Block diagrams of the analyzer and synthesizer are shown in Figures 22 and 23.

All of the experiments in this section have the following common features:

1. The adaptive modes i = 1, 2, and 4 correspond to frame intervals of duration FI = 10i ms.

2. The sets of frame decisions $\{i_r\}$ and pitch decisions $\{p_r\}$ were obtained by hand in the preliminary analysis, and provided as inputs to the vocoder simulation.

3. Minimum-phase synthesis is employed.

4. Linear interpolation is accomplished in the convolution operation between adjacent voiced frames.

Experiments 4 through 7 were conducted with test Sentence 1 as the input. Experiments 8 and 9 used Sentences 2 and 3 respectively.

Experiment 4

This experiment has three parts. Parts 1 and 2 correspond to Parts 1 and 2 of Experiment 2. In Part 3 the adaptive processor used windows of duration $D_i = 12.8i$ ms for adaptive modes i = 1, 2, and 4. The cepstrum was truncated to $K_i = 10i$ coefficients without quantization. The experiment is summarized as follows:

	Adaptive Mode i	Window Duration D (ms)	Frame Interval FI (ms)	Number of Cepstrum Coefficients per Frame K	Number of Coefficients "Transmitted" per second
Part 1		25.6	20	24	1200
Part 2		12.8	10	24	2400
Part 3	1	12.8	10	10	1000
	2	25.6	20	20	
	4	51.2	40	40	

As in Experiment 2, Part 2 was judged to be the best of the three. Parts 1 and 3 were judged to be of comparable "quality."

Experiment 5

This experiment has three parts. In Parts 1 and 2, the conventional processor was employed, with windows of duration 51.2 and 25.6 ms, and frame intervals of 20 and 10 ms. The adaptive processor in Part 3 was operated only in modes i = 1 and 2. The mode 4 frames of Experiment 4 were replaced by mode 2 frames. In all three parts the cepstrum coefficients were quantized to 6 bits. The experiment is summarized as follows:

	Adaptive Mode i	Window Duration D (ms)	Frame Interval FI (ms)	Number of Coefficients K	Estimated Bit Rate (b/s)
Part 1		51.2	20	24	4800
Part 2		25.6	10	24	9600
Part 3	1	12.8	10	10	4000
	2	25.6	20	20	

The adaptive processor of Part 3 was judged to yield the best synthesized speech. A distinct improvement in "quality" was noticeable in the adaptive results. Part 3 was compared in a listening test to the original sentence. Only a very slight degradation in "quality" was noticeable in the vocoded speech.

Experiment 6

The two parts of this experiment compare the conventional and adaptive vocoder. In Part 1 the conventional vocoder was operated with a 40 ms window, updating 26 cepstrum coefficients every 20 ms. The experiment is summarized as follows:

	Adaptive Mode i	Window Duration D (ms)	Frame Interval FI (ms)	Number of Coefficients K	Approximate Bit Rate BR (b/s)
Part 1		40	20	26	6800
Part 2	1	12.8	10	10	5700
	2	25.6	20	20	

In both parts the cepstrum coefficients were quantized to 7 bits. The approximate bit rates were calculated as follows. The range of the quantization was the maximum range encountered in the cepstrum (-1 to 8). An examination of the envelope of the cepstrum revealed that coefficients 2 through 5 occupy only half the range of the first coefficient, and thus, may be coded in one less bit. Similarly, coefficients above number 5 occupy only half the range of 2 through 5, and thus, require one less bit. For a conventional vocoder employing K coefficients quantized uni-

formly to Q bits and a frame interval FI ms, the approximate bit rate (BR) is

$$BR = [Q + 4(Q - 1) + (K - 5)(Q - 2)]\frac{10^3}{FI} \text{ b/s} \qquad (5\text{-}1)$$

The bit rate of the adaptive processor is dominated by the shortest frame, since transmitting fewer coefficients requires a greater average number of bits per coefficient. All the adaptive experiments in this chapter employ $K_1 = 10$ coefficients and $FI_1 = 10$ ms, so the approximate bit rate is

$$BR = [1 + Q + 4(Q - 1) + 5(Q - 2)]\frac{10^3}{10} \qquad (5\text{-}2)$$

$$= 1000Q - 1300 \qquad \text{b/s}$$

The added term (unity) in the bracket of (5-2) accounts for the 1 bit per vocoder frame that must be transmitted to the synthesizer to signal the adaptive mode, i = 1 or 2.

In the listening test Part 2 was judged to be superior in "quality" to Part 1.

Experiment 7

In this experiment the adaptive vocoder was operated with cepstrum quantization to 7, 6, 5, and 4 bits. Otherwise the adaptive processor of this experiment is identical to that of Part 2 of Experiment 6. The experiment is summarized as follows:

Part	Quantization to Q bits	Approximate Bit Rate BR (b/s)
1	7	5700
2	6	4700
3	5	3700
4	4	2700

The "quality" of Parts 1, 2, and 3 was judged to be the same, while a slight degradation became noticeable in Part 4. The degradation was similar to that one observes on a fading long-haul high-frequency radio channel. The synthesized speech of Part 3 was quite natural, and was judged to retain speaker recognition ability. The Part 3 result was compared to the original speech in a listening test. Only a slight degradation in "quality" was noticeable in the synthesized speech.

The waveforms encountered in the adaptive vocoder of Part 3 are plotted in Figure 35. The frame illustrated is r = 61, which begins 620 ms into the sentence. This mode 2 frame represents a segment of the phoneme /I/ in "gift is." The measured pitch for the frame is 5.3 ms.

Narrow- and wide-band Sonagrams of the synthesized speech of Part 3 of this experiment are pictured in Figure 36(a). The Sonagrams of the original speech were pictured in Figure 9(a) of Chapter II.

Experiment 8

In this experiment the adaptive vocoder processed test Sentence 2. Window durations of 10 and 20 ms were used in modes 1 and 2. The cepstrum coefficients were quantized to 5 bits. The experiment is summarized as follows:

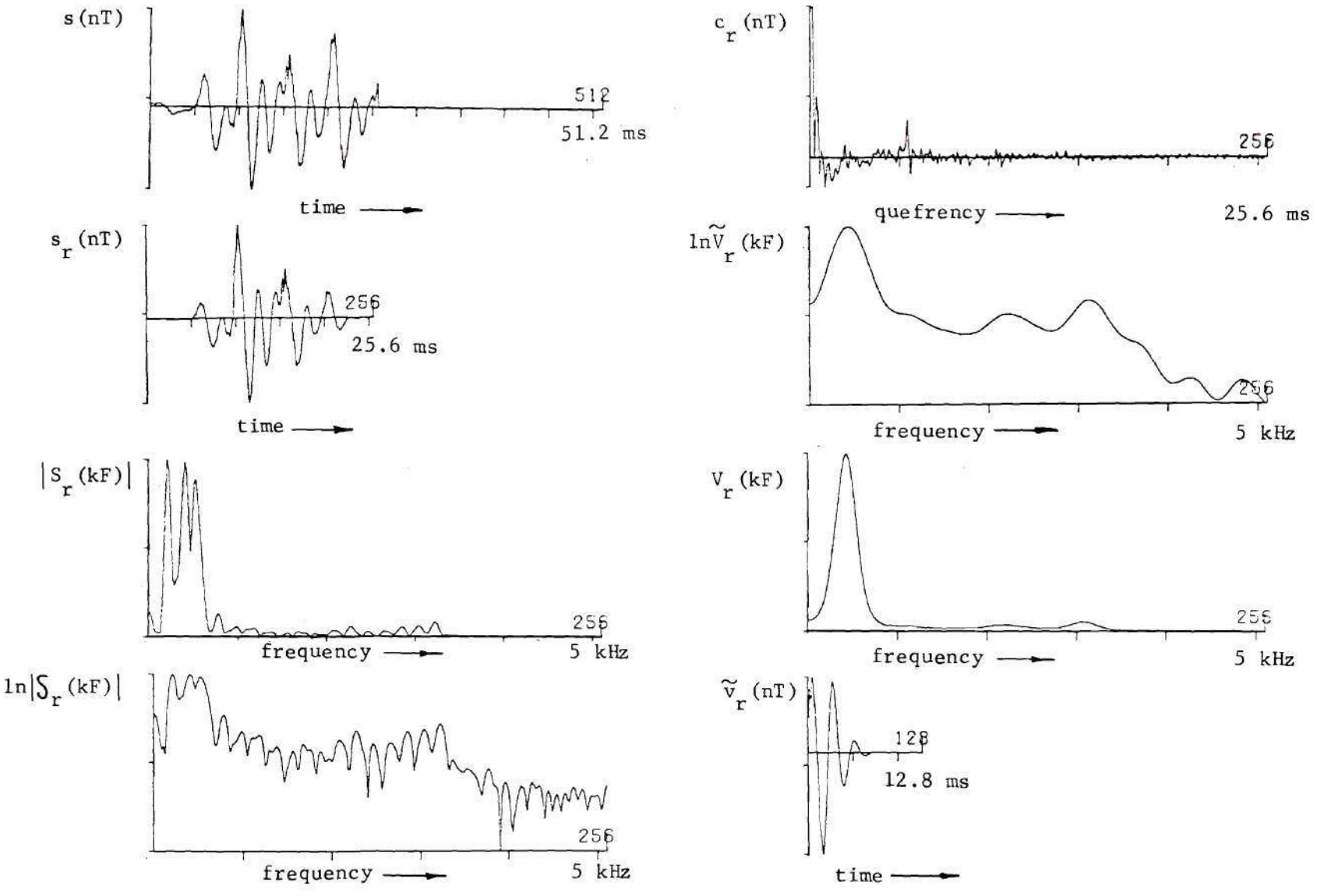


Figure 35. Vocoder Waveforms - Part 3 of Experiment 7

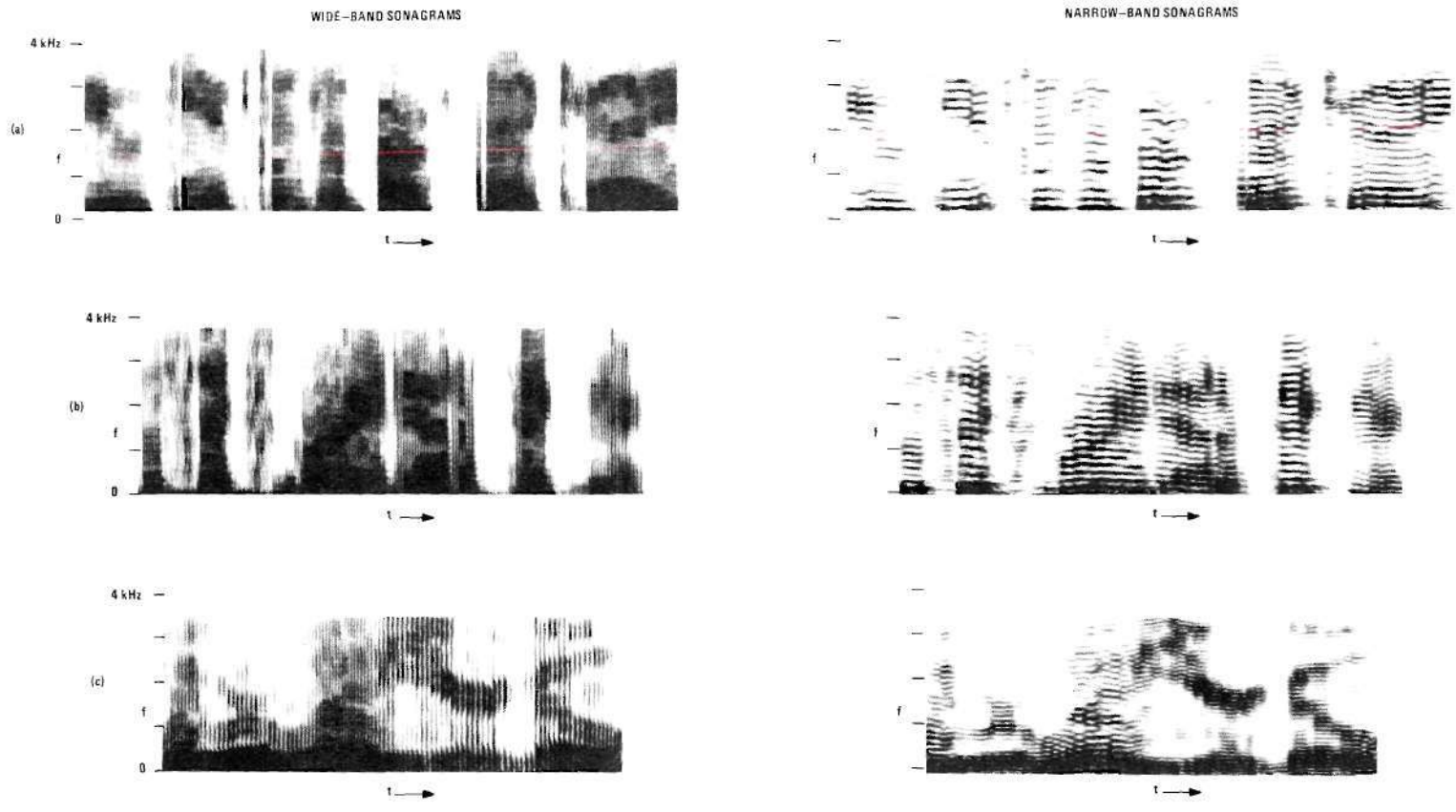


Figure 36. Sonagrams of Vocoded Speech.
(a) Sentence 1 -- Part 3 of Experiment 7.
(b) Sentence 2 -- Experiment 8.
(c) Sentence 3 -- Experiment 9.

Adaptive Mode i	Window Duration D (ms)	Frame Interval FI (ms)	Number of Coefficients K	Approximate Bit Rate BR (b/s)
1	10	10	10	3700
2	20	20	20	

For the listening test the vocoded result of this experiment was compared to the original sentence. The synthetic speech quality was reasonably good, but not quite as good as that obtained using 12.8 and 25.6 ms windows in Part 3 of Experiment 7 (in which Sentence 1 was the input). The intelligibility was judged to be very high, and speaker recognition ability retained.

Narrow- and wide-band Sonagrams of the vocoded speech of this experiment are pictured in Figure 36(b), which may be compared to Sonagrams of the original speech in Figure 9(b).

Experiment 9

In this experiment Sentence 3 was processed by the adaptive vocoder of Experiment 8. Since Sentence 3 has no rapid transitions, at least none rapid enough to warrant use of the short window, the entire sentence was processed in mode 2. Thus, in this "adaptive" run, the vocoder operated exactly as a conventional vocoder with D = 20 ms, FI = 20 ms, K = 20, and Q = 5 bits. The approximate bit rate was BR = 3700 b/s.

As in Experiment 8, the vocoded result of this experiment was reasonably good, but not as good as the result obtained previously with Sentence 1.

Narrow- and wide-band Sonagrams of the vocoded speech of this experiment are pictured in Figure 36(c), which may be compared to Sonagrams of the original speech in Figure 9(c).

Summary

The results obtained with the adaptive vocoder suggest that the adaptive scheme has potential for reducing the required spectrum data rate while retaining intelligibility, naturalness and speaker recognition properties. The spectrum bit rate was reduced from 6800 b/s to 3700 b/s without noticeable degradation in "quality" as judged in informal subjective listening tests. Some degradation became noticeable when the data rate was reduced to 2700 b/s.

Conclusion

The experimental results obtained with the adaptive homomorphic vocoder indicate that the adaptive strategy has some potential for improving the "quality" of vocoded speech. But this is a tentative conclusion based on only a few experiments with three test sentences. A much more exhaustive series of experiments will be required in order to judge the effectiveness of the adaptive strategy.

CHAPTER VI

RECOMMENDATIONS FOR FURTHER WORK

Introduction

Six areas of further investigation are proposed in this chapter. Additional experimental study is clearly required to judge the potential of the adaptive strategy. A specific program of further investigation is outlined.

Two variations of the conventional homomorphic vocoder are described, analysis with gaps and "fepstrum" analysis. Preliminary experimental results suggest that both variations may have potential.

A technique called adaptive bandpass coding is motivated by an example. The scheme offers a way to retain more of the high frequency "information" of unvoiced sounds.

An interesting approach to the deconvolution of voiced speech is described. The approach requires solution of a system of linear equations and an independent means of pitch extraction.

The chapter concludes with a brief discussion of the potential savings in storage which may be possible in voice response units using speech coded from adaptive spectrum analysis.

Evaluation of the Adaptive Vocoder

The experimental results reported in Chapter V provide only a tentative indication of the potential of the adaptive vocoder strategy. A

more exhaustive series of experiments will be required to judge the effectiveness of the adaptive strategy.

An outline of proposed further investigation follows:

1. Improve the vocoder simulation system. Paper tape is a very fragile medium for communicating digitized speech between computers. The major improvement required in the simulation system described in Chapter IV is to implement a digital magnetic tape link between the UNIVAC 1108 and the small computer used for D/A conversion. Such a link will significantly improve the versatility of the system.

2. Conduct a thorough evaluation of the time-frequency resolution properties of the conventional homomorphic vocoder. Vary the analysis resolution rectangle between extreme limits and judge the effect on the "quality" of the resulting synthesized speech. Such experiments will provide results to guide the choice of time-frequency cells for later experiments.

3. Examine the effects of cepstrum quantization. Study the tradeoff between the number (K) of cepstrum coefficients transmitted (i.e., the frequency resolution) and the quantization of the coefficients (e.g., to Q_i bits for the i^{th} coefficient). While holding the spectrum data rate approximately constant ($\sum_{i=1}^{K} Q_i$ = constant), vary K and choose the Q_i to accomplish uniform quantization within the envelope of possible cepstrum sequences. Repeating this experiment for several bit rates will yield an experimental criterion for choosing K and $\{Q_i\}$ for a given bit rate.

4. Evaluate the adaptive homomorphic vocoder. Incorporate the results of previous experiments into the design of a conventional homomorphic vocoder to serve as a reference for comparison. Operate the adaptive vocoder over a wide range of bit rates and judge the resulting synthesized speech.

The investigation outlined above should provide sufficient experimental evidence to make a positive conclusion about the effectiveness of the adaptive approach to speech analysis-synthesis. It is anticipated that the adaptive processor will allow vocoder operation at significantly lower bit rates without loss of "quality."

Analysis with Gaps

One result of Experiment 3 (described in Chapter V) was that satisfactory vocoded speech resulted from analysis with a 12.8 ms window and a 20 ms frame interval. This result suggests an interesting approach to reducing the complexity of the homomorphic vocoder.

Since the analysis window duration controls the size of the FFT required, and thus, the complexity of the vocoder itself, the use of a "short" window function would substantially reduce the complexity of the homomorphic vocoder. Further investigation is required to discover whether or not analysis with gaps will allow satisfactory synthesized speech for a wide range of talkers and test sentences.

A Variation on the Cepstrum -- the "Fepstrum"

Early in the research effort, consideration was given to possible improvement in vocoded speech that might be gained by including phase

information in the spectrum coding. The results reported by Flanagan and Golden [38] with the so-called phase vocoder were motivating, as was the deconvolution of speech Oppenheim performed using the complex cepstrum [17].

The "fepstrum" is defined as the cepstrum of the even part of a real, causal, time-limited sequence of length $\frac{N}{2}$-1. The fepstrum appears to retain more "phase information" than does the conventional cepstrum, since the input sequence may be recovered from the fepstrum if a binary phase sequence of length $\frac{N}{2}$+1 is known.

The fepstrum is discussed in Appendix A. The results of a preliminary experiment suggest that the fepstrum may be useful in speech processing.

Adaptive Bandpass Coding

One topic which has not been emphasized in this thesis is the passband of speech that should be coded. Telephone circuits typically transmit only a 300-3500 Hz range of the speech spectrum, and very high intelligibility results. But the spectrum of some phonemes, especially the fricative and stop consonants, has considerable energy above 3500 Hz.

The vocoder simulation reported in this thesis processed input speech which was low-pass filtered at 4 kHz and sampled at 10 kHz. Thus, the spectrum sections coded by the cepstrum have range 0-5 kHz, with little energy above 4 kHz.

Suppose one had the task of implementing a vocoder which retained an 8 kHz range of the speech spectrum. The homomorphic vocoder offers

a simple way to achieve this goal without increasing the channel data rate. Notice that by truncating the spectrum in a given vocoder frame one may achieve the same effect as reducing the sampling rate of the input speech.

We illustrate the adaptive bandpass coding approach with the following example. Suppose a homomorphic vocoder is to be implemented which retains an 8 kHz range, which employs a D = 40 ms analysis window, and which transmits 30 cepstrum coefficients for each 20 ms frame. Further, suppose the input speech is low-pass filtered at 8 kHz and sampled at 20 kHz, so T = .05 ms.

The spectrum of the voiced sounds of speech (even those which contain an unvoiced component) is dominated by the energy below 4 kHz. Conversely, unvoiced sounds have considerable energy above 4 kHz. Thus, we are motivated to adapt the pass band represented by the cepstrum coding between 4 kHz for voiced frames and 8 kHz for unvoiced frames.

Let the subscripts 1 and 2 denote the voiced and unvoiced conditions respectively. The input speech is windowed, resulting in an 800 point sequence $s_r(nT)$. Computing the DFT with an N = 1000 point FFT yields the amplitude spectrum $|S_r(kF)|$, which has 501 samples spaced $F = \frac{1}{NT} = 20$ Hz over the range 0-10 kHz (the remaining 499 samples are the even symmetric image of the first 501).

In the voiced mode i = 1 we truncate $|S_r(kF)|$ to 5 kHz (by retaining only the first 251 samples) and compute the cepstrum with an $N_1 = 500$ point FFT (where the FFT input is composed of the even extension of the first 251 samples of $\ln |S_r(kF)|$). The resulting cepstrum $c_1(n2T)$ has

samples spaced $\frac{1}{N_1 F} = 2T$ sec. Truncating the cepstrum to K = 30 coefficients (3 ms) for transmission, we retain approximately 200 Hz frequency resolution in the spectrum coding.

In the unvoiced mode i = 2 the full 10 kHz range is retained in $|S_r(kF)|$, and the cepstrum $c_2(nT)$ computed with an $N_2 = 1000$ point FFT and truncated to K = 30 coefficients (1.5 ms) for transmission. Thus, approximately 400 Hz resolution is retained in the spectrum coding. The synthesizer operates in a manner complementary to the analyzer. That is, in mode i = 1, $\ln \tilde{V}(kF)$ is computed with an $N_1 = 500$ point FFT. Before the final FFT operation the (complex) sequence $\tilde{V}(kF)$ (which has "conjugate symmetry," $\tilde{V}((N_1 - k)F) = \tilde{V}^*(kF)$) is augmented by N_1-1 zeroes. The resulting impulse response v(nT) is sampled at the desired rate, the interpolated samples resulting from the augmentation by zeroes. In mode i = 2 the synthesizer operates in the conventional manner, with $N_2 = 1000$.

Notice that the adaptive operation is keyed to the v/uv condition, so that no additional bits need be transmitted to signal the adaptive mode.

An alternative implementation, which would produce equivalent results, is the following. FFT's with N = 1000 are used in all transform operations. In mode i = 1, the sequence $|S_r(kF)|$ is truncated by zeroing the central 499 samples (which effectively discards components above 5 kHz). The resulting cepstrum $c_1(nT)$ is sampled at twice the necessary rate. Thus, only the (K = 30) samples of even index (n = 0, 2, ..., 58) need be transmitted. At the synthesizer, the cepstrum samples are alternated with zeroes before the FFT.

Rather than merely truncating the spectrum, one should consider windowing the spectrum with a "smooth" function such as the Hanning window to gain improved "quefrency resolution" in the cepstrum. The same argument applies to truncating the cepstrum. Some improvement in vocoded speech "quality" may result.

The adaptive bandpass coding scheme demonstrates again the versatility of the homomorphic vocoder.

One may speculate that adaptive bandpass coding will yield improved "quality" vocoded speech. In addition, the scheme might be useful in the data-reduction coding of any wide-band signal source which has short-term narrow-band characteristics.

Should the adaptive bandpass coding approach prove useful, it might logically be combined with the adaptive spectrum analysis strategy described in this thesis.

A New Approach to the Deconvolution of Speech

An assumption inherent to the application of homomorphic deconvolution to speech is that the window function be approximately constant over the duration of the vocal tract impulse response function $v(t)$. Consideration of this assumption leads to an approach to the deconvolution of speech which appears to compensate for the effect of the window function and yield accuracy limited only by numerical constraints. The approach is outlined in Appendix B.

The scheme requires an independent pitch detector. The complex samples of the short-time DFT, $S_r(kF)$, are used in a system of linear equations

$$S_r(kF) = \sum_n V(\frac{n}{T_o})\ W(kF - \frac{n}{T_o}) \qquad (6\text{-}1)$$

A subset of this system of equations is solved for $V(\frac{n}{T_o})$, the vocal tract system function sampled at the pitch frequency $1/T_o$.

One may speculate that this approach is potentially useful in the deconvolution of speech.

Voice Answerback Applications

Adaptive spectrum analysis appears to have potential application to voice answerback applications. In the communications (vocoder) problem, a practical constraint is that the reduced data representation of speech be generated at a uniform bit rate. Such a constraint does not apply to voice answerback.

Suppose the adaptive vocoder were operated in 4 modes, with the added mode (i = 0) reserved to describe a 10 ms silent frame. The analysis is summarized as follows:

Mode i	Frame Interval FI (ms)	Number of Cepstrum Coefficients K	Number of Bits Stored
0	10	0	2
1	10	10	38
2	20	20	68
4	40	30	98

We assume quantization to Q = 5 bits. If we assume the 4 modes to be equally likely, then the average bit rate for spectrum information is

approximately 2580 b/s, compared to approximately 3800 b/s for an adaptive vocoded representation of equivalent quality.

Summary

In this chapter six areas of further research are described. Additional experimental study is clearly required to judge the effectiveness of the adaptive spectrum analysis strategy.

Two variations on the conventional homomorphic vocoder, analysis with gaps and "fepstrum" analysis, appear to have potential usefulness.

Adaptive action of the homomorphic vocoder in another dimension, i.e., the frequency range to be coded, was described. Adaptive bandpass coding seems to offer a simple way to improve the "high frequency response" of the vocoder during unvoiced sounds.

A novel approach to the deconvolution of speech was described which appears to have potential for "high quality" deconvolution of voiced speech. The scheme involves solution of a linear system of equations, and requires an independent pitch detector.

Finally, the potential application of adaptive spectrum analysis to voice answerback was discussed briefly. An hypothetical example illustrates that adaptively coded spectrum information results in a reduced storage requirement.

CHAPTER VII

SUMMARY

The investigation reported in this thesis was focused on improving the spectrum information employed in speech analysis and synthesis systems. Modern vocoders use a fixed time-frequency compromise in the analysis and subsequent synthesis of speech.

The phonemes of speech display a wide range of time-frequency properties, due to the extremes in the articulatory dynamics of speech production. The vocoder is based on the simplified model of speech production, an essentially "stationary" model. The relative validity of the model may be improved by matching the duration of the analysis window function to intervals of speech which are indeed "stationary." These observations motivated the design and experimental evaluation of a vocoder which adapts its time-frequency resolution properties to match the relative stationarity of different segments of input speech.

The homomorphic vocoder was selected as a test platform to evaluate the adaptive spectrum analysis strategy. The homomorphic vocoder was a natural choice for the simulation because its time and frequency properties may be readily manipulated, time resolution by the duration of the analysis window function, and frequency resolution by the number of cepstrum coefficients transmitted.

An adaptive homomorphic vocoder was designed and a simulation sys-

tem implemented on a large-scale digital computer. A series of experiments was conducted with the adaptive vocoder. In one experiment the spectrum data rate was reduced from 6800 to 3700 b/s without noticeable loss of "quality" of the synthesized speech. The tentative conclusion drawn from the experimental results is that the adaptive strategy appears to have potential for substantially reducing vocoder data rates, while maintaining intelligibility, speaker recognition, and naturalness properties.

APPENDIX A

THE "FEPSTRUM"

Let us define the "fepstrum" (for "funny cepstrum") to be the cepstrum of the even part $s_e(nT)$ of a real, causal time-limited sequence $s(nT)$,

$$s(nT) = 0 \qquad n \leq 0,\ n \geq \frac{N}{2} \tag{A-1}$$

$$s_e(nT) = \frac{s(nT) + s((N - n)T)}{2} \qquad (0 \leq n \leq N-1)$$

The fepstrum is

$$\underline{c}(nT) = DFT^{-1}\left\{ \ln \left| DFT\{s_e(nT)\} \right| \right\} \tag{A-2}$$

$$= \frac{1}{N} \sum_{k=0}^{N-1} \ln \left| \sum_{m=0}^{N-1} s_e(mT) e^{-j2\pi km/N} \right| \qquad e^{+j2\pi kn/N}$$

Substituting (A-1) into (A-2) and manipulating yields

$$\underline{c}(nT) = \frac{1}{N} \sum_{k=0}^{N-1} \ln \left| \sum_{m=0}^{\frac{N}{2}-1} s(mT) \cos(2\pi km/N) \right| e^{+j2\pi kn/N} \tag{A-3}$$

$$= \frac{1}{N} \sum_{k=0}^{\frac{N}{2}-1} \ln \left| \sum_{m=0}^{\frac{N}{2}-1} s(mT) \cos(2\pi km/N) \right| \cdot \cos(2\pi kn/N)$$

where the second line follows because the inner summation is even in k about $k = N/2$.

The DFT of $s_e(nT)$ is identical to the real part of the DFT of the sequence formed by augmenting $s(nT)$ with $N/2$ zeroes. Notice that $S_e(kF) = \text{DFT}\ \{s_e(nT)\}$ is a real sequence of length N, even in k about $k = N/2$, so it is completely specified by its first $\frac{N}{2} + 1$ samples. If we ignore numerical error, $S_e(kF)$ may be inverted to obtain $s_e(nT)$, from which $s(nT)$ may be recovered exactly since $s(nT) = 2\, S_e(nT)$ for $0 \leqq n \leqq \frac{N}{2} - 1$.

The $|\cdot|$ operation in (A-3) discards some "phase information," but, since $S_e(kF)$ is real and even, $S(nF)$ may be recovered from $|S_e(kF)|$ if the phase function $\Phi(kF) = \text{sgn}[S_e(kF)]$ is known for $0 \leqq k \leqq \frac{N}{2} + 1$. Notice that $\Phi(kF)$ is a binary sequence, completely specified by $\frac{N}{2} + 1$ bits. The fepstrum $\underline{c}(nT)$ is a real sequence of length N, even in n about $n = \frac{N}{2}$, so it is described by its first $\frac{N}{2} + 1$ samples. If the binary phase sequence $\Phi(kF)$ is known, the input sequence $s(nT)$ may be recovered exactly from the fepstrum.

Contrast the invertability of the fepstrum with that of the conventional cepstrum:

$$c(nT) = \frac{1}{N} \sum_{k=0}^{\frac{N}{2}-1} \ln \left| \sum_{m=0}^{\frac{N}{2}-1} s(mT)\, e^{-j2\pi km/(N/2)} \right| \cdot e^{+j2\pi kn/(N/2)} \qquad \text{(A-4)}$$

The inner sum is $S(k\frac{2}{NT})$, the $\frac{N}{2}$ point DFT of $s(nT)$. Since $s(nT)$ is real, $S(k\frac{2}{NT})$ has even real part and odd imaginary part. Thus, $S(k\frac{2}{NT})$ is

described by $\frac{N}{2}$ real numbers. The $|\cdot|$ operation in (A-4) may only be inverted if the phase function $\theta(k\frac{2}{NT})$ is known, where $S(k\frac{2}{NT}) = |S(k\frac{2}{NT})|e^{j\theta(k\frac{2}{NT})}$. θ is odd in k about $k = \frac{N}{4}$, so it may be described by $\frac{N}{4} - 1$ numbers. The cepstrum c(nT) is real and even about $n = \frac{N}{4}$. c(nT) has length $\frac{N}{2}$, and may be described by $\frac{N}{4} + 1$ samples.

Notice that the fepstrum contains twice as many samples as the cepstrum, but the fepstrum may be inverted to obtain s(nT) if only a binary phase sequence of $\frac{N}{2} + 1$ bits is known. Conversely, inverting the cepstrum requires knowledge of a $\frac{N}{4} - 1$ point sequence. One may argue that the fepstrum contains substantially more "phase information" than does the cepstrum.

Using the fepstrum in place of the cepstrum in the homomorphic vocoder leads to a synthesized vocal tract system function $\tilde{V}(kF)$, the magnitude of which is a smoothed version of $|S_e(k\frac{2}{NT})|$.

One may expect the envelope of the real part of $S_e(k\frac{2}{NT})$ to be very similar to the envelope of $S(\frac{k}{NT})$. In addition, S_e appears to retain more frequency resolution than S, since a window of twice the duration is used in computing S_e.

In several early vocoder analysis-synthesis runs, fepstrum analysis was combined with conventional minimum phase synthesis. Plots of the resulting synthesized spectrum $|\tilde{V}(kF)|$ were essentially equivalent to those obtained using cepstrum analysis. Speech synthesized using the two analysis techniques was generated using monotone pitch excitation. No difference in speech quality was discernable.

APPENDIX B

DECONVOLUTION OF SPEECH -- A NEW APPROACH

This appendix outlines a new approach to the deconvolution of speech which appears to offer the ability to recover the vocal tract spectrum, including phase information, with accuracy limited only by numerical constraints. The scheme requires an independent pitch detector. The vocal tract spectrum V(kF) is obtained as the solution of a system of linear equations.

We motivate the approach by examining the effect of the window function in the homomorphic vocoder. A stationary segment of voiced speech is assumed to be generated as the convolution of an excitation signal e(t), which is modeled as a train of impulses with spacing T_o equal to the pitch period, with a vocal tract impulse response function v(t).

The input speech is windowed by a function w(t) with duration D, short enough to guarantee that T_o and v(t) are stationary. The result is

$$s(t) = [e(t) \circledast v(t)]\, w(t) \qquad \text{(B-1)}$$

where

$$e(t) = \sum_n \delta(t - nT_o)$$

$$\longleftrightarrow E(f) = \sum_n \delta\left(f - \frac{n}{T_o}\right)$$

The assumption that (B-1) may be approximated as

$$s(t) \cong [e(t)\ w(t)] \otimes v(t) \tag{B-2}$$

is valid to the extent that $w(nT_o) = w((n + 1)T_o)$. In other words, the assumption depends on how closely the windowed speech of (B-1) may be approximated by a summation of the convolution components $v(t - nT_o)$, each of which has been weighted by a constant $w(nT_o)$. That is:

$$w(t) \sum_n v(t - nT_o) \cong \sum_n w(nT_o)v(t - nT_o) \tag{B-3}$$

Another view of the approximation may be gained in the frequency domain. Transforming (B-1), we obtain

$$S(f) = [E(f)V(f)] \otimes W(f) \tag{B-4}$$

$$= \int_{-\infty}^{\infty} W(\mu)\ E(f - \mu)\ V(f - \mu)\ d\mu$$

which is a product of (complex) spectra to the extent that $W(\mu)$ is impulsive in frequency. Substituting E(f) we obtain

$$S(f) = \sum_n V(\frac{n}{T_0})\ W(f - \frac{n}{T_o}) \tag{B-5}$$

Similarly, from (B-2) we obtain

$$S(f) \cong [E(f) \otimes W(f)]\ V(f) \tag{B-6}$$

$$\cong V(f) \sum_{n} W(f - \frac{n}{T_o})$$

so the assumption that (B-1) may be replaced by (B-2) is valid if $W(\frac{n}{T_o}) \cong 0$ for $n \neq 0$, which is approximately true if

$$B_L = \frac{4}{D} < \frac{1}{T_o} \tag{B-7}$$

$$D > 4T_o$$

Notice that the summation in (B-5) resembles a sampling theorem expansion of S(f) in translated versions of the "kernel" $W(f - \frac{n}{T_o})$. The approximation in (B-6) may be interpreted as follows. The samples of S(f) which coincide with pitch harmonics at $f = \frac{n}{T_o}$ are approximately equal to the samples of the vocal tract system function

$$S(\frac{n}{T_o}) \cong V(\frac{n}{T_o}) \tag{B-8}$$

Thus, the envelope of S(f) is an approximation to the vocal tract amplitude spectrum $|V(f)|$.

The homomorphic vocoder separates the vocal tract function $|V(f)|$, with accuracy limited by the degree of validity of the approximation (B-2). Notice that homomorphic deconvolution does not compensate for the effect of the window function.

Suppose we sample the exact expression in (B-5) at frequencies $f = kF$, to obtain

$$S(kF) = \sum_{n} V(\frac{n}{T_o})\ W(kF - \frac{n}{T_o}) \qquad \text{(B-9)}$$

and express the result in matrix notation

$$S_k = W_{kn}\ V_n \qquad \text{(B-10)}$$

where $S_k = \{S(kF)\}$ is a column vector with K rows, $V_n = \{V(\frac{n}{T_o})\}$ is a column vector with N rows, and $W_{kn} = \{W(kF - \frac{n}{T_o})\}$ is a rectangular $K \times N$ array. If V(f) is essentially band limited to f_1 Hz, then $N = [f_1 T_o] + 1$. Similarly $K = [f_1/F] + 1$.

Equation (B-10) is a linear system of K equations in N unknowns V_n. We speculate that an $N \times N$ subset of this system may be solved to yield V_n, the complex samples of the vocal tract system function. Suppose F is selected to be $1/T_o$. Then the kernel W_{kn} is dominated by its diagonal terms (to the extent that $D > T_o$) and has conjugate symmetry about its diagonal. If w(t) is assumed to be centered at the origin, then W_{kn} is real and even about its diagonal.

Notice that the deconvolved V(kF) retains phase information, but not absolute phase, since the e(t) used in (B-1) has no phase reference. But notice that an arbitrary delay term introduced in e(t) (e.g., $e(t - T_1)$) causes a complex exponential factor in E(f) (e.g., $e^{-j2\pi fT_1}$) which may be

lumped into V(kF). Transforming V(kF) yields $v(nT - T_1)$, so that phase coherence may be retained from frame to frame with this technique.

If (B-10) may indeed be solved for V_n, the vocal tract function obtained appears to include no ill effects of the window function. Thus, we speculate that very short windows, with duration on the order of one pitch period may yield high quality results.

The need to solve a system of linear equations in each frame appears to limit the potential usefulness of this scheme to non real-time application.

BIBLIOGRAPHY

1. J. L. Flanagan, Speech Analysis, Synthesis and Perception, Academic Press, Inc., New York, 1965.

2. J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic Voices for Computers," IEEE Spectrum, vol. 7, pp. 22-45, October 1970.

3. H. Dudley, "The Carrier Nature of Speech," The Bell System Technical Journal, vol. 19, pp. 495-515, October 1940.

4. G. Fant, Acoustic Theory of Speech Production, Mouton & Co., The Hague, Netherlands, 1960.

5. A. Papoulis, The Fourier Integral and Its Applications, McGraw-Hill Book Co., Inc., New York, 1962.

6. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Computation of Complex Fourier Series," Mathematics of Computation, vol. 19, pp. 297-301, April 1965.

7. A. V. Oppenheim, "Speech Spectrograms Using the Fast Fourier Transform," IEEE Spectrum, vol. 7, pp. 57-62, August 1970.

8. B. Gold and C. M. Rader, Digital Processing of Signals, McGraw-Hill Book Co., Inc., New York, 1969.

9. P. Mermelstein, "Computer-Generated Spectogram Displays for On-Line Speech Research," IEEE Transactions on Audio and Electroacoustics, vol. AU-19, pp. 44-47, March 1971.

10. B. Gold and C. M. Rader, "The Channel Vocoder," IEEE Transactions on Audio and Electroacoustics, vol. AU-15, pp. 148-161, December 1967.

11. T. Bially and W. M. Anderson, "A Digital Channel Vocoder," IEEE Transactions on Communications Technology, vol. COM-18, pp. 435-442, August 1970.

12. R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," Journal of the Acoustical Society of America, Vol. 47, pp. 634-648, February 1970.

BIBLIOGRAPHY (Continued)

13. C. J. Weinstein and A. V. Oppenheim, "Predictive Coding in a Homomorphic Vocoder," Massachusetts Institute of Technology, Lincoln Laboratory Preprint, DS-9872, January 1971.

14. A. V. Oppenheim, "Superposition in a Class of Non-linear Systems," Massachusetts Institute of Technology, Research Laboratory of Electronics, Technical Report 432, March 31, 1965.

15. A. V. Oppenheim, "Generalized Linear Filtering," in Digital Processing of Signals, B. Gold and C. M. Rader, McGraw-Hill Book Co., Inc., New York, 1969, chapter 8.

16. A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "The Nonlinear Filtering of Multiplied and Convolved Signals," Proceedings of the IEEE, vol. 56, pp. 1264-1291, August 1968.

17. A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Transactions on Audio and Electroacoustics, vol. AU-16, pp. 221-226, June 1968.

18. A. V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering," Journal of the Acoustical Society of America, vol. 45, pp. 458-465, February 1969.

19. A. M. Noll, "Cepstrum Pitch Determination," Journal of the Acoustical Society of America, vol. 41, pp. 293-309, February 1967.

20. B. Bogert, M. Healy, and J. Tukey, "The Quefrency Alanysis of Time Series for Echoes," in Proceedings of the Symposium on Time Series Analysis, M. Rosenblatt, Ed., John Wiley and Sons, Inc., New York, 1963, chapter 15.

21. H. J. Manley, H. Shaffer, C. R. Howard, and J. O'Brien, "All-Digital Vocoder Techniques," Technical Report AFAL-TR-69-302, Air Force Avionics Laboratory, Wright-Patterson Air Force Base, Ohio, December 1969.

22. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II," The Bell System Technical Journal, vol. 40, pp. 65-84, January 1961.

23. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - III: The Dimension of the Space of Essentially Time- and Band-Limited Signals," The Bell System Technical Journal, vol. 41, pp. 1295-1366, July 1962.

BIBLIOGRAPHY (Continued)

24. P. M. Woodward, Probability and Information Theory, with Applications to Radar, McGraw-Hill Book Co., Inc., New York, 1953.

25. R. B. Blackman and J. W. Tukey, The Measurement of Power Spectra, Dover Publications, Inc., New York, 1958.

26. D. C. Rife and G. A. Vincent, "Use of the Discrete Fourier Transform in the Measurement of Frequencies and Levels of Tones," The Bell System Technical Journal, vol. 49, pp. 197-228, February 1970.

27. D. Gabor, "Theory of Communications," Journal of the IEE (London), vol. 93, pp. 429-457, November 1946.

28. A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill Book Co., Inc., New York, 1965.

29. C. W. Helstrom, "An Expansion of a Signal in Gaussian Elementary Signals," IEEE Transactions on Information Theory, vol. IT-12, pp. 81-82, January 1966.

30. R. M. Lerner, "Representation of Signals," in Lectures on Communication System Theory, E. J. Baghdady, Ed., McGraw-Hill Book Co., Inc., New York, 1961, chapter 10.

31. L. K. Montgomery and I. S. Reed, "A Generalization of the Gabor-Helstrom Transform," IEEE Transactions on Information Theory, vol. IT-13, pp. 344-345, April 1967.

32. A. W. Rihaczek, "Signal Energy Distribution in Time and Frequency," IEEE Transactions on Information Theory, vol. IT-14, pp. 369-374, May 1968.

33. C. I. Malme, "Detectability of Small Irregularities in a Broadband Noise Spectrum," Massachusetts Institute of Technology, Research Laboratory of Electronics, Quarterly Report, January 1959.

34. M. Lecours and J. J. Sparkes, "Adaptive Spectral Analysis for Speech-Sound Recognition," IEEE Transactions on Audio and Electroacoustics, vol. AU-16, pp. 523-525, December 1968.

35. B. O. Pyron and F. R. Williamson, "Study and Analysis of Speech Parameters and Bandwidth Compression Techniques," Final Report, Contract DA-49-092-ARO-156, Army Research Office, Washington, D. C., June 1967.

BIBLIOGRAPHY (Concluded)

36. W. D. Voiers, "The Present State of Digital Vocoding Technique: A Diagnostic Evaluation," IEEE Transactions on Audio and Electroacoustics, vol. AU-16, pp. 275-279, June 1968.

37. M. L. Uhrich, "Fast Fourier Transforms Without Sorting," IEEE Transactions on Audio and Electroacoustics, vol. AU-17, pp. 170-172, June 1969.

38. J. L. Flanagan and R. M. Golden, "Phase Vocoder," The Bell System Technical Journal, vol. 45, pp. 1493-1509, November 1966.

VITA

Jack Curtis Hammett, Jr., son of Jack Curtis and Barbara Poe Hammett, was born in Little Rock, Arkansas, on December 29, 1939. He married Carol Anne Raines of Little Rock, Arkansas in December, 1960.

After attending public school in Little Rock, he entered the University of Arkansas in 1957, and was graduated in 1961 with a B.S. in Electrical Engineering. Upon graduation, he was commissioned into the United States Army Signal Corps and served as platoon leader, communications staff officer, and company commander in tactical units in the United States and on Okinawa. In 1965-1966 he served with the 173d Airborne Brigade in the Republic of Viet Nam. He attended the Signal Officer Career Course at Fort Monmouth, New Jersey, and served on the staff of the Signal School. He entered the Georgia Institute of Technology in June, 1967, and received the M.S.E.E. degree in June, 1969, at which time he entered the Ph.D. program.